金融行业

数据安全专报

1 月报
2021 年

安全月报

政策解读 | 技术实践 | 安全观察 | 解决方案

绿盟科技金融事业部出品

政策解读

绿盟专家谈热点 |
《数据安全法（草案）》向社会征求意见

技术实践

个人金融信息保护视角下的
脱敏效果评估研究与实践

大数据下的隐私攻防：数据脱敏后的
隐私攻击与风险评估

数据匿名化：隐私合规下，企业打开
数据主动权的正确方式？

透过隐私合规，看数据安全技术
发展趋势

数据安全治理体系之解析

NSFOCUS

贴身服务 加油干

绿盟科技城商行信息安全解决方案

无缝衔接 | 密切配合

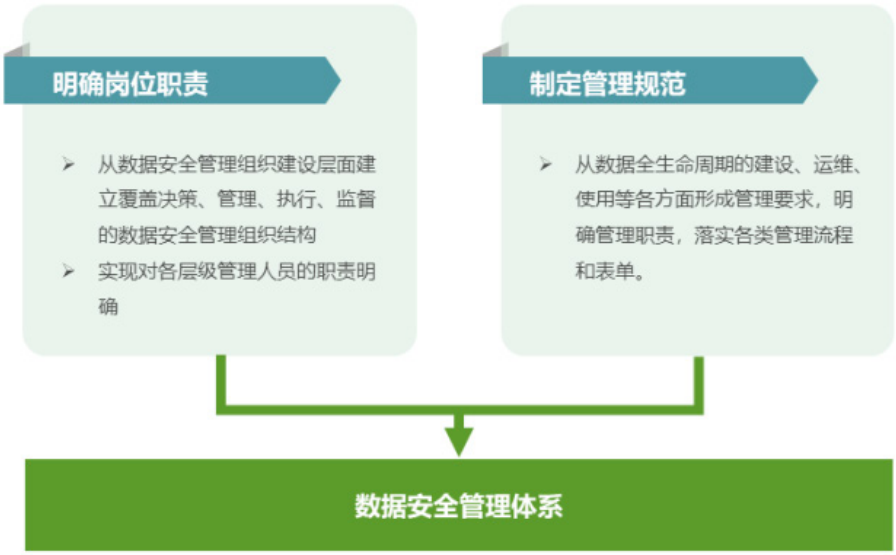
THE EXPERT
BEHIND GIANTS
巨人背后的专家

多年以来，绿盟科技致力于安全攻防的研究，
为金融、政府、运营商、能源、互联网以及教育、医疗等行业用户，提供具有核心竞争力的安全产品及解决方案，帮助客户实现业务的安全顺畅运行。
在这些巨人的背后，他们是备受信赖的专家。

NSFOCUS绿盟科技

本|期|看|点

P4 绿盟专家谈热点 | 《数据安全法（草案）》向社会征求意见



P42 大数据下的隐私攻防：数据脱敏后的隐私攻击与风险评估

Figure 1: Number of breaches reported by 9/30 each year

Figure 2: Number of records lost (in millions) by 9/30 each year

安全月报

2021年第1期

绿盟科技金融事业部

安全月报在线阅读

绿盟科技官方微信

目录 CONTENTS

政策解读

技术实践

安全观察

解决方案

NSFOCUS

政策解读

绿盟专家谈热点 | 《数据安全法（草案）》向社会征求意见

绿盟科技　数据安全特工队

业界呼声颇高的数据安全法草案，在2020年6月28日-30日举行的十三届全国人大常委会第二十次会议迎来初次审评，这也代表着我国数据安全保护从此有了法律依据。

背景介绍

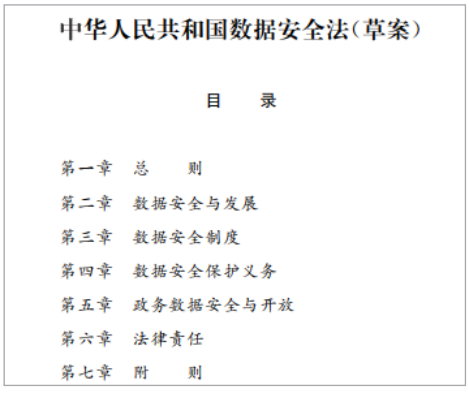
随着信息技术与经济社会的交汇融合，大数据技术及应用蓬勃发展，大数据数量和价值的快速攀升，大数据时代的数据安全也面临着巨大挑战。

从近年来发生的数据泄露事件可以看出，数据安全问题已经影响到国家安全发展，关系到公众利益，与公民权益密切相关。与此同时，欧盟、美国、日本等国家相继出台数据保护法案。

在全球各国围绕数据的争夺和博弈不断深化的背景下，国家领导人对数据安全也高度重视。

法律介绍

2018年9月7日，十三届全国人大常委会公布立法规划，《中华人民共和国数据安全法》位于第一类项目：条件比较成熟、任期内拟提请审议的法律草案，由委员长会议负责起草，2020年6月28日-30日举行的十三届全国人大常委会第二十次会议迎来初次审议。



中华人民共和国数据安全法（草案）

目　录

《数据安全法》和《网络安全法》作为《国家安全法》的配套法规，是国家整体安全观的组成部分，在适用范围和保护职责上各有侧重、互相补充，共同建筑网络安全和数据安全。

《数据安全法》的诞生标志了数据安全上升到国家安全层面。

主要内容

数据安全法是总体国家安全观框架下，国家安全法律体系的重要组成部分。该法律在网络安全法的基础上，进一步明确了数据安全相关者的保护义务与职责，并与《数据安全管理办法（征求意见稿）》相互照应。《数据安全法》的诞生，标志着数据安全上升到国家安全层面，意义重大。数据安全法共七章五十一条。

《数据安全法（草案）》主要内容包括：

按照总体国家安全观的要求，确立数据安全保护管理各项基本制度，提升国家数据安全保障能力，有效应对数据这一非传统领域的国家安全风险与挑战，切实维护国家主权、安全和发展利益；

坚持安全与发展并重，规定支持、促进数据安全与发展的措施，提升数据安全治理和数据开发利用水平，促进以数据为关键要素的数字经济发展；

立足数据安全工作实际，着力解决数据安全领域突出问题，落实数据活动主体的安全保护义务与责任，切实维护公民、组织的合法权益；

适应电子政务发展的需要，建立政务数据安全管理制度和开放利用规则，大力推进政务数据资源开放和开发利用。

关注焦点

绿盟科技数据安全咨询专家在第一时间对《数据安全法（草案）》进行解读和分析，总结出7点《数据安全法（草案）》的关注焦点。

01. 数据安全责任制，落实数据全生命周期管控责任

建立数据安全组织架构，明确岗位职责，制定对应的全流程管理规范、制度、流程等。

设计健全的组织架构是数据安全管理工作的基础，从数据安全管理建设角度建立决策层、管理层、执行层、监督层等多方面、跨部门有效协同的机制与制度，明确各数据安全岗位职责，实现对各层级管理、执行人员的责任落实。同时，从数据安全生命周期的各个阶段形成管理要求，包括方针、规章制度、管理标准、管理规范、管理流程执行表单等。

02. 数据分类分级，实现企业数据安全建设第一步

建立数据资产管理机制，明确保护对象及策略。

数据分类分级服务是基于法律法规以及业务需求确定组织内部的数据分类分级方法，帮助组织理清数据资产，对生成或收集的数据进行分类标识，并以数据分类为基础，采用规范、明确的方法区分数据的重要性和敏感度差异进行分级管理，确定数据重要性或敏感度，针对性地采取适当、合理的管理措施和安全防护措施，形成科学、规范的数据资产管理与保护机制。

03. 发现企业数据安全隐患，降低数据安全风险

利用风险评估手段识别发现企业的数据安全风险，协助企业进行整改，提升企业数据安全建设水平。

数据安全风险评估是以数据为中心，识别发现数据环境以及数据行为是否存在侵害数据或侵犯数据主体权利风险的过程。数据安全风险评估依据数据分类分级保护的需求，符合企业实际情况的方式梳理风险检查项，全面发现数据全生命周期存在的问题，并按照高、中、低3个风险等级进行管理，有效提升数据安全性，让数据安全风险可控。

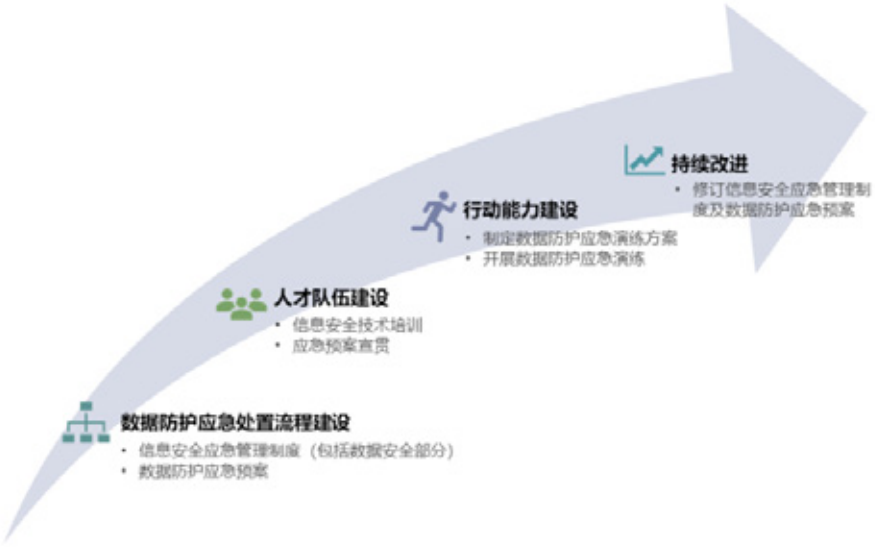
04. 识别全流程数据活动，落实数据安全控制措施

梳理数据全生命周期，制定相对应的安全要求，对各风险点进行提示，包含可落地执行的机制等。

与传统信息系统安全不同，数据是流动的，数据安全管控措施的落实是一个以数据为中心的动态过程。明确管控措施的落实策略，通过对业务实现中数据的流动方式进行分析，根据数据流识别出业务中的各种数据活动。只有识别出数据活动，才能够准确识别数据在在流动中处于信息系统的具体环境，进而在具体环境下落实相应的数据生命周期各环节安全管控措施。

05. 建立数据安全事件应急响应机制

建立数据防护应急预案，明确数据安全事件的应急方针、政策，应急组织结构及相关应急职责。



建立信息安全应急管理体系，包括数据防护应急管理体系。建立应急工作领导小组、应急工作管理小组、应急执行小组与应急工作联络小组，承担指挥、组织、决策、通知、实施等工作职责。制定数据防护专项应急预案，对数据安全事件进行明确定义。发生安全事件时，迅速调用专项应急预案，快速有效解决问题，恢复系统平稳运行。定期开展数据防护应急管理体系宣贯工作，定期组织开展应急演练和应急管理体系优化工作。

06. 组织开展数据安全培训教育

组织开展数据安全专业培训，提升企事业单位数据安全保护意识，加强数据安全人员专业能力提升。

开展数据安全培训教育工作，包括全员安全意识培训，数据安全基础知识培训，数据安全应急预案培训与数据安全专项能力培训。安全意识培训中，重点对人员安全行为进行普及，防止人员的无意识行为导致数据安全事件的发生。数据安全基础知识培训中，需要了解企业内部数据安全组织架构形式，并对数据安全的全生命周期环节中的要求点有所了解。数据安全应急预案培训中，明确数据防护应急流程，包括数据安全事件的分级、安全事件的上报流程、事件处理相关方式方法及流程，事后内部报告改进与对外消息传播流程等。数据安全专项培养中，不仅要求人员了解数据安全的基本要求，还要对其衍生出的个人信息安全、隐私安全、敏感数据安全等进行专项培训。

07. 聚焦政务数据安全和开放，保证开放共享平台安全

建立政务数据全流程管理规范、制度、流程等，明确政务数据分级管控流程。

坚实的数据安全建设是政务数据的开放共享的前提，建立健全地区政务数据开放共享政策体系，让政务数据在安全的状态下充分流通，并采取妥善的管理与技术措施保障政务数据开放共享平台的安全，真正发挥政务数据的价值，促进当地经济发展，提升公民生活质量。

总结

通过对数据安全法的解读，以及绿盟科技现阶段对数据安全的独特见解，我们认为，未来企业应明确数据安全组织职责，设立数据安全官角色（参考GDPR中DPO角色），为数据安全工作落地执行提供强有力的人力资源上的保障。同时，应强调管理和技术双管齐下，保障数据始终处在安全的环境下充分流通，在确保安全的同时让数据产生最大的价值。从管理角度看，数据安全的管理制度、保障措施、岗位职责等需要依托数据分级分类进行编制。从技术实现角度看，不同类别和级别的数据需采取不同的安全防护措施，从而实现安全保护与实际业务需求的有效协同。

浅析数据安全与隐私保护之法规

在大数据时代背景下，AI和大数据技术给我们的生活带来了巨大的便利和效率；然而在此过程中，数据滥用、数据窃取、隐私泄露以及“大数据杀熟”等数据安全问题呈徒增和爆发趋势。

在这样背景下，全球各个国家纷纷颁布相关法规，对数据安全与隐私保护相关问题进行严格的规范与引导。如欧盟保护个人数据的《General Data Protection Regulation》(简称《GDPR》)；美国的《California Consumer Privacy Act》(简称《CCPA》)；中国实施的《中华人民共和国网络安全法》（通常简称《网安法》）。法规作为数据安全治理和建设的顶层指导，研究这些法规，一方面有助于更好地理解安全场景与需求，进而有利于将安全技术实际的落地与应用；另一方面提前研究，通过积极开展数据安全治理与防护，可以提前规避企业由于数据不合规带来的法律风险和处罚。如最近一年中有两个典型的案例：Google旗下的子公司Alphabet，在欧洲因个人数据的处理违反欧盟GDPR法规，被罚5000万欧元；Facebook泄露门事件——8700万用户信息泄露，面临美国50亿美元的天价罚单（相当于公司一年利润收入的20%多）。这两个事件足以说明数据安全建设对于一个企业来说，多么重要且迫切。

数据安全问题广受社会各界关注：包括学术界，研究隐私保护与数据挖掘等关键技术一直是近年来的热点方向；工业界，寻找具体的可落地的数据安全解决方案是企业重要的战略方向。笔者作为该领域的其中一员，认为现阶段的数据问题不同于数据安全问题，应结合大数据时代背景进行讨论和分析：数据的最终目标应该是使用和开发，而不是一味强调隐私保护和数据保密性。因此可以称为“大数据时代下的数据安全”，笔者将该主题在该系列中拆分成三个部分进行介绍：相关法规、场景与技术、实践体系，本文是这个系列中的第一篇：相关法规篇。除整理外，笔者也会兴致使然，发表几点思考和解读，希望能起到抛砖引玉的作用，最终与各位专家共同探讨。

1 国外相关法规

研究国外法规，可以对比且了解全球立法和执行趋势。另一方面，国外的法规对国内的企业在该外国境内的数据处理以及数据的跨境传输，同样有法律影响和效力。本文以欧盟的《通用数据保护条例》和美国的《加州隐私保护法》的法律框架为例，给出几个点的简单分析与介绍。

1.1 欧盟《通用数据保护条例》

欧盟于2018年5月25日正式实施了《通用数据保护条例》（《General

Data Protection Regulation》,简称《GDPR》)[1]，是一项保护欧盟公民个人隐私和数据的法律，其适用范围包括欧盟成员国境内企业的个人数据、也包括欧盟境外企业处理欧盟公民的个人数据。

图1 欧盟个人数据保护法规《GDPR》

《GDPR》由11章99个条款组成，是一项的“大而全”的个人数据保护框架，因此非常值得深入研究。由于篇幅所限，这里仅列3点进行分析：

①“个人数据”的定义?

GDPR：第四条，“个人数据”是关于一个已识别或者可能识别的自然人（即数据主体）的任何信息。一个可能识别的自然人是能够直接或间接识别的人，尤其是借助标识符例如该自然人的姓名、身份证号码、位置数据、在线标识符，或者自然人的一个或多个特定因素的组合，比如物理、心理、遗传、精神、经济、文化属性或社会身份等。

解读：“个人数据”是隐私保护相关法律重要基础。GDPR采用宽泛的“个人数据”定义，尽可能包含所有可能的、与自然人相关的“个人数据”，这些数据都受到GDPR的监管和保护。

详细地说，其定义的“个人数据”包括两类：一类是“已识别”，比如包括“详细住址”和“姓名”的信息是可以唯一识别（或者定位）到具体的“自然人”；另一类是“可能识别”，在GDPR的前言26段进一步解释说，“可能识别”是在合理和可能的条件下（比如成本和时间）进行。它包括两个子类别：一个标识符，典型的是身份证号、手机号等，它并不能直接识别到“自然人”，但若可以掌握额外的数据库、比如身份证/手机号映射“详细住址”和“姓名”，它是可能被识别的。另一个是“自然人”其他碎片化属性的组合（一个或者多个以上），比如性别、邮编、身高、体重、从事职业，这些属性的数据可以经过各种组合，如在某一个班级场景中，性别和身高有可能识别到唯一的“自然人”。

这个定义表明GDPR保护的数据主体范围非常之广泛和庞大，不但包括容易枚举的个人数据，姓名、身份证号、地址等。也包括一些模糊拓展的边界，例如网络的属性cookie，IP地址码, Mac地址码，以及生物识别数据，指纹、虹膜等这些某些特殊的场景下，仍然可能关联或者定位到唯一的“自然人”[2]。这个场景让人联想到一些企业正在使用的“大数据画像/千人千面技术”。采集用户静态属性(如年龄和性别)和行为信息(比如浏览、点击和收藏等)，当收集的用户信息（特征）足够多，足以将这个人唯一“区分”出来。那么这些静态或者动态将成为个人数据，理应是GDPR监管的范畴。宽泛的定义，可以最大限度保护好“自然人”的隐私，规避一些“擦边球”的场景。但企业如何识别在复杂的业务环境中去识别这些数据、如何更好地处理和保护这些数据。这些无疑给企业的合规策略、流程、以及技术的实施带来巨大的挑战，同时意味着在企业需要在数据安全建设上投入更多的经费和支持。

② 用户如何行使 GDPR 赋予的“被遗忘权”？

GDPR：GDPR赋予了用户(数据主体)知情权、访问权、修正权、删除权（被遗忘权）、限制处理权（反对权）、可携带权、拒绝权等7项基本权利，其中删除权是一项引人注目的用户“特权”之一。即在第十七条中，在个人数据已不再是数据控制者和处理者的收集和处理目的等6种情况下，赋予了用户删除权。

解读：官网提供有两个有趣的例子[3]。Example 1：假如你加入了一个社交网站，但过了一段时间，你决定要离开这个社交网站。那么你可以行使“被遗忘权”，即要求公司删除属于你的个人数据。 Example 2：当你用你的名字使用搜索引擎搜索时，出现一个很久以前，与你有关的的债务偿还协议，但现在已无关紧要。对于一个普通人，他有权利要求删除这些网络信息。用户除了以上的权利，数据主体还拥有知情权、访问权、修正权、限制处理权（反对权）、可携带权、拒绝权等基本权利。但笔者认为，“被遗忘权”是GDPR赋予用户非常大的一项权利。我们每个人平均一年要注册10多个网站/APP，但很多网站/APP是低频访问或者后续一直不用的状态。但用户的个人数据却被互联网公司收集和存储和处理，用户感觉到“个人数据扩散不可控”。如实说，笔者对一些公司网站的“注册”有种恐惧感，例如个人数据被贩卖第三方从而收到骚扰电话或广告邮件的轰炸、另外一些“小网站”被黑客攻击造成数据泄露的概率也更高。若这些网站/APP提供一键删除注册信息的功能（举例好像题跑偏了，那就假如我国后续的立法也赋予了用户类似的权利），作为用户，肯定乐于行使该项权利。

图2 产品功能的畅想：一键删除低频访问的APP/网站的“注册信息”

③ GDPR 如何惩罚违规的企业？

GDPR：第83条给出犯规罚款的最高值。即可被处以最高2000万欧元的行政罚款，或对企业以最高占上一财政年度全球总营业额4%的行政罚款，取两者最高值。

解读：这是欧盟境内企业或者与欧盟的数据相关的境外企业最关心的。GDPR给的惩罚力度相比其他国家、地区非常严厉。从2018年5月执行的一年以来，GDPR开出多个罚单。其中，Google处罚5000万是最大一张罚单。预测

在不久将来，欧盟各个国家将开出更多的罚单。这条法规是迫使相关企业投入更多资金，部署数据安全产品，马上行动，进行数据安全治理与建设直接源动力。

1.2 美国《加州消费者隐私法》

美国已有多个州先在数据安全与隐私保护进行了立法，其中最著名的要数2018年6月加州通过《加州消费者隐私法案》（《California Consumer Privacy Act》，简称《CCPA》）[4]。该法案被称为美国“最严厉和最全面的个人隐私保护法案”，将于2020年1月1日生效。

①“个人信息”的定义？

CCPA：“个人信息”系指直接或间接地识别、关系到、描述、能够相关联或可合理地连结到特定消费者或家庭的信息。

解读：CCPA称为“个人信息”，其实和“个人数据”是同一个概念。受GDPR的影响，CCPA同样采用了类似宽泛的定义。根据定义，罗列了出11种个人信息类别，包括姓名、驾照等信息，也有互联网IP标识符、标识符。有特点的是，把反映消费者偏好、特征、心理倾向、偏好、倾向、行为、态度、智力、能力和资质的画像也列入了个人信息范畴。比GDPR更加广泛的“个人信息”范畴给企业个人敏感数据的梳理、治理带来了巨大的工作量和挑战。

② 用户如何行使 CCPA 赋予的“访问权”？

CCPA：在CCPA的1798.100中，企业从可验证消费者处收到要求访问个人信息的请求后，应立即采取措施向消费者免费披露和提供本节所要求的个人信息。个人信息的提供可通过信件或电子方式，如果以电子方式提供，信息应以便携方式提供并且在技术可行限度内采用易于使用的形式。

解读：CCPA也赋予了消费者知情权、访问权、删除权、限制处理权和拒绝权等权利。CCPA访问权的特色在于回复的“及时性”、反馈的“便携式”——邮件发送等形式，这比GDPR赋予的“访问权”更加具体。当用户需要查看或确认采集信息，无疑这条法规提供了极大的便利。但反过来说，给企业造成了一定负担。

③ CCPA 如何惩罚违规的企业？

CCPA：在1798.155中，对于法规企业，每次违规行为可能要承担高达7,500美元的民事罚款。另外在1798.145中，对于违规企业，为每个消费者每次事件赔偿不少于100美元且不超过750美元的赔偿金，以数额较大者为准。

解读：罚款最有特色的地方，考虑到消费者的损失且量化为影响的消费者数量，将赔偿每一个消费者100-750元的赔偿金，这与GDPR的罚款机制不同。若一个企业违规涉及的消费者信息有100w条，那么它至少要承担1亿美元的罚款。

2 国内相关法规

我国在数据安全与个人信息上目前涉及的法规有《中华人民共和国刑法》（以下简称《刑法》）、《最高人民法院、最高人民检察院关于办理侵犯公民个人信息刑事案件适用法律若干问题的解释》（以下简称《若干问题的解释》）、《中华人民共和国网络安全法》、《电信和互联网用户个人信息保护规定》。下面以《中华人民共和国网络安全法》和处在征求意见稿阶段的《数据安全管理办法》为例，进行简要的分析和介绍。

2.1《中华人民共和国网络安全法》

我国于2017年6月1日正式实施《中华人民共和国网络安全法》（通常简称《网安法》）[5]。《网安法》是我国首部全面规范网络空间安全管理方面问题的基础性法律，包含的内容十分丰富，一共包括7章79条，包含网络运行安全、关键信息基础设施

的运行安全、网络信息安全等内容。值得关注的是，《网安法》在数据（包括个人信息）安全与保护上也有诸多规定，例如第四十至四十五条。

① 数据安全与数据利用技术发展的关系？

原文：第十八条，国家鼓励开发网络数据安全保护和利用技术，促进公共数据资源开放，推动技术创新和经济社会发展。

解读：这条法规表明我国对数据持开放和发展态度，并没有一味强调数据安全保护，而是强调数据最终目的是利用与开放，同步发展数据安全保护与利用技术，有利于社技术创新和社会发展。

②“个人信息”的定义？

原文：是指以电子或者其他方式记录的能够单独或者与其他信息结合识别自然人个人身份的各种信息，包括但不限于自然人的姓名、出生日期、身份证件号码、个人生物识别信息、住址、电话号码等。

解读：以“识别”为核心，包括直接识别的如身份证号、姓名等，间接识别，如性别属性，由于结合出生年月、地址识别的个人身份，因此也是个人信息。相比GDPR和CCPA来说，我国罗列的个人信息范畴不大，并不包括由个人关联的信息，比如用户的行为/习惯、购买的IoT设备等识别性不高的信息，这对于国内企业治理是一件好事，缩小了“个人信息”的范围，降低敏感信息分类分级的成本。然而，在颁布的《信息安全技术个人信息安全规范》推荐标准中，重新拓展了个人信息的范畴，给出个人信息判定，不仅包括“识别”，还包括“关联”，从个人到信息。这条规定饱受一定的争议，因为“自然人”每时每刻都在生产各种各样的“个人信息”，比如在公共电脑或设备上操作留下的日志，某一个人在纸上写了一个字，这些都成为了个人信息。若按照推荐标准，对个人信息/敏感数据进行梳理和分类，那么无疑让企业将陷入“无穷无尽”的困境。建议同时也期待后续的《数据安全管理办法》、《个人信息保护法》以及出台的相关标准，对“个人信息”定义作进一步清晰和明确的界定，便于企业落地实施与合规。

③“个人信息”开放 / 共享的出路在哪？

原文：第四十二条规定，网络运营者不得泄露、篡改、毁损其收集的个人信息；未经被收集者同意，不得向他人提供个人信息。但是，经过处理无法识别特定个人且不能复原的除外。

解读：数据的价值在于流动，在于再次利用与挖掘。这对于“以人为本”的

个人信息同样适用，“个人信息”大数据最终目标应该是造福人类、服务社会。但未经过处理的“个人信息”，在数据共享和开放过程中风险不容忽视，被不法分子利用，如通过精确的个人信息进行高级的电信诈骗，造成“徐玉玉案”的悲剧。因此，共享前，进行特殊处理与风险评估，显得十分必要。假如一批个人信息经过处理，已经“无法识别特定个人且不能复原”，那么即使公开，大家也只能获得一些统计和分析信息，无法得到定向到“自然人”，那么该记录里面的实际关联的“自然人”，自然受到利益损失或者威胁。为了达到这个处理的目的，研究和开发相应的技术是必要的。

2.2《数据安全管理办法》（征求意见稿）

2019年5月28日国家互联网信息办公室发布《数据安全管理办法》（征求意见稿）[6]。在《网安法》的指导下，该法规对数据安全作进一步做了详细的规定和约束。（由于处于征求意见稿阶段，因此不作深入解读）。它明法规的管理范围是中华人民共和国境内利用网络开展数据收集、存储、传输、处理、使用等活动（第二条），数据安全分为个人信息和重要数据安全（第一条）。征求意见稿中包括数据收集、数据处理使用和数据安全监督管理等内容。数据处理内容中值得关注的是，第二十三条，网络运营者利用用户数据和算法推送新闻信息、商业广告等（以下简称“定向推送”），应当以明显方式标明“定推”字样，为用户提供停止接收定向推送信息的功能；用户选择停止接收定向推送信息时，应当停止推送，并删除已经收集的设备识别码等用户数据和个人信息。这条对当前火热的用户画像技术对这一行为进行了严格的约束，提升了用户的体验和个人数据的安全。

3 小结

在大数据时代，数据安全与隐私问题显得越来越严峻。为了应对挑战，全球掀起了数据安全与个人信息的立法热潮。欧盟的 “大而全”且十分严格的《GDPR》作为一个风向标，或多或少影响了其他国家的立法，特别是美国加州的《CCPA》。国外的法规，特别是处罚力度之大，迫使企业尽快进行数据治理与建设，另一方面也很好地保护了用户的切身利益。我国已经颁发的《网安法》虽然没有给出违反“个人信息”规定的企业相应量化处罚与罚款措施，但在《刑法》和《若干问题解释》有相关规定，如出售个人信息50条可入罪等量刑场景。

随着后续一些法规的颁布和实行，如《数据安全管理办法》、《个人信息保护法》以及一些行业法规，以及一些配套的标准的实现，如《个人信息去标识化指南》、《个人信息安全影响评估指南》等等，未来几年中，我国数据安全相关法规-标准建设将趋于体系化和成熟。相应地，法规的处罚机制也将更加清晰与明朗。

图3 国内数据安全相关法规-标准体系

实际上，数据安全相关的立法并不是完全限制住大数据产业的利用与开发；相反，它可以引导大数据相关产业朝向健康、安全且持久的正确方向发展。在各个国家法规中，都可以或多或少找到一些数据利用和开发的出口，包括欧盟的《GDPR》，我国的《网安法》给出了相应的灵活解释——特别是对于匿名化数据进行豁免，不受法规约束，即这类数据可以像非敏感数据一样开发、共享与使用。

参考资料

1.《General Data Protection Regulation》, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv: OJ.L_.2016.119.01.0001.01.ENG&toc=OJ:L:2016:119:TOC

2. https://ec.europa.eu/info/law/law-topic/data-protection/reform/what-personal-data_en

3. https://ec.europa.eu/info/law/law-topic/data-protection/reform/rights-citizens/my-rights/can-i-ask-company-delete-my-personal-data_en

4.《California Consumer Privacy Act》, https://cal-privacy.com/

5.《中华人民共和国网络安全法》http://xxzx.mca.gov.cn/article/wlaqf2017/wjjd/201705/20170500891068.shtml

6.《数据安全管理办法》（征求意见稿）http://www.moj.gov.cn/news/content/2019-05/28/zlk_235861.html

《金融数据安全 数据安全分级指南》解读

引言

进入大数据时代，随着金融信息化的蓬勃发展，众多金融基础设施、业务系统、应用程序产生的数据也呈指数级增长，而随之而来的数据安全问题也日益严峻。由于金融行业的特殊性，金融数据安全问题不仅仅影响金融业自身稳定发展，也会对金融市场稳定、公众社会秩序，甚至国家安全造成威胁。习近平总书记也多次强调，金融是国家重要的核心竞争力，金融安全是国家安全的重要组成部分。因此，金融机构亟需有针对性的金融数据安全标准或规范来指导金融机构合理使用和管理金融数据，并围绕金融数据建立完善的数据安全防护体系。

2020年9月23日，中国人民银行正式印发《金融数据安全 数据安全分级指南》(JR/T 0197-2020) 金融行业标准(以下简称《标准》)，并自发布之日正式开始实施。该《标准》由中国人民银行提出，全国金融标准化技术委员会归口管理，起草单位包括金融监管单位、安全检测机构、金融业机构、金融科技公司以及大学等数十家单位。《标准》不仅适用于指导金融业机构开展数据安全分级工作，并可为第三方评估机构等单位开展数据安全检查与评估等相关工作提供参考。笔者将针对《标准》的核心内容进行梳理解读，并结合绿盟科技在数据安全领域的技术积累给出应对建议。

《金融数据安全 数据安全分级指南》核心内容

《标准》明确指出了金融数据安全分级的目标、原则和范围，数据安全定级的工作流程，并给出了金融业机构典型数据定级规则供实践参考。

◆ 定级目标：

数据安全定级旨在对数据资产进行全面梳理并确立适当的数据安全分级，是金融业机构实施有效数据分级管理的必要前提和基础。数据分级管理是建立统一、完善的数据生命周期安全保护框架的基础工作，能够为金融业机构制定有针对性的数据安全管控措施提供支撑。

◆ 定级原则：

数据安全定级遵循以下原则：

原则	详情
合法合规性原则	满足国家法律法规及行业主管部门有关规定
可执行性原则	定级规则避免过于复杂，确保可行性
时效性原则	数据安全级别具有一定的有效期限，按需及时调整级别

原则	详情
自主性原则	金融业机构在本标准框架下按需自主确定数据安全级别
差异性原则	应根据数据的类型、敏感程序等差异，划分不同数据安全级别
客观性原则	数据定级规则是客观且可校验的

◆ 定级范围：

安全定级工作所涉及的金融数据包括但不限于：

1. 提供金融产品或服务过程中直接或间接采集的数据。

2. 金融业机构信息系统内生成和存储的数据，包括业务数据、经营管理数据等。

3. 金融业机构内部办公网络与办公设备（终端）中产生、交换、归档的电子数据。

4. 金融业机构原纸质文件经过扫描或其他电子化手段形成的电子数据。

5. 其他宜进行分级的金融数据。

注意：金融数据安全定级过程中，未经电子化的金融数据，依据档案文件等有关管理规范执行；涉及国家秘密的金融数据，依据国家有关法律法规执行，不在本标准规定的范围之内。证券行业数据安全分级工作可参照JR/T 0158-2018执行。

◆ 定级要素：

数据的安全性（保密性、完整性、可用性）遭到破坏后可能造成的影响（如可能造成的危害、损失或潜在风险等），是确定数据安全级别的重要判断依据，主要考虑影响对象与影响程度两个要素。

定级要素	内容简述
影响对象	包括国家安全、公众权益、个人隐私、企业合法权益。
影响程度	从高到低划分为严重损害、一般损害、轻微损害和无损害。

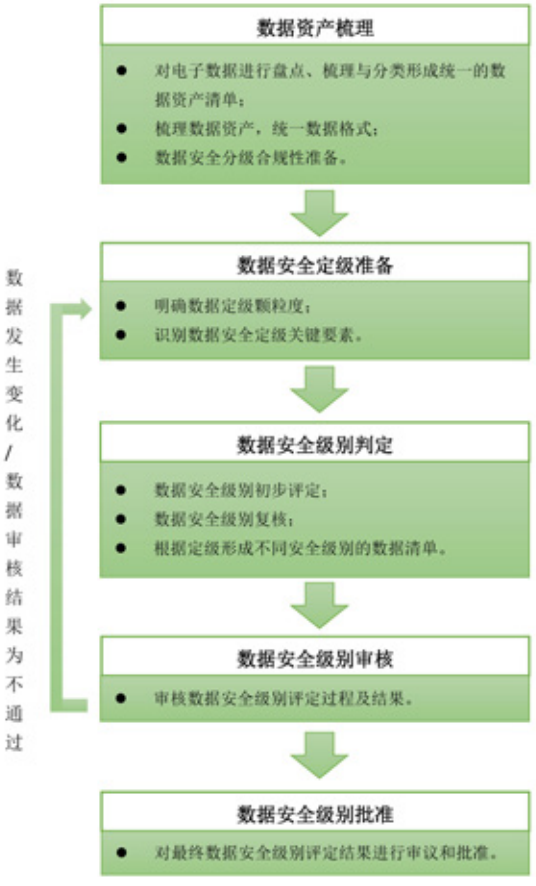
◆ 安全影响评估：

安全影响评估宜综合考虑数据类型、数据内容、数据规模、数据来源、机构职能和业务特点等因素，对数据安全性（保密性、完整性、可用性）遭受破坏后所造成的影响进行评估。评估过程中，根据实际情况识别各项安全性在影响评定中的优先级，分别进行保密性、完整性及可用性评估，并综合考虑保密性、完整性及可用性的评估结果，形成最终安全影响评估。

◆ 定级规则：

最低安全级别参考	数据定级要素	
	影响对象	影响程度
5	国家安全	严重损害 / 一般损害 / 轻微损害
5	公众权益	严重损害
4	公众权益	一般损害
4	个人隐私	严重损害
4	企业合法权益	严重损害
3	公众权益	轻微损害
3	个人隐私	一般损害
3	企业合法权益	一般损害

◆ 定级过程：



金融业典型数据类型及其建议划分的最低安全级别，参见《标准》附录A。

◆ 级别变更：

数据安全定级完成后，出现下列情形之一时，宜对相关数据的安全级别进行变更：

1	数据内容发生变化，导致原有数据的安全级别不适用变化后的数据。
2	数据内容未发生变化，但因数据时效性、数据规模、数据使用场景、数据加工处理方式等发生变化，导致原定的数据安全级别不再适用。
3	因数据汇聚融合，导致原有数据安全级别不再适用汇聚融合后的数据。
4	因国家或行业主管部门要求，导致原定的数据安全级别不再适用。
5	需要对数据安全级别进行变更的其他情形。

导致数据发生升降级的主要技术手段有数据脱敏、删除关键字段、汇聚融合等（详情参见《标准》附录B）：

措施	安全级别调整
汇聚融合	3 级升至 4 级
生产数据脱敏后用于金融业机构内部业务经营或管理工作	3 级降至 2 级
汇聚融合，特定机构特定时间或事件后信息具有高安全等级	2 级升至 4 级
脱敏，从数据中去除能够直接定位到个人金融信息主体的内容，删除涉及商业秘密的内容等，特定时间或事件后信息失去原有敏感性	4 级降至 2 级

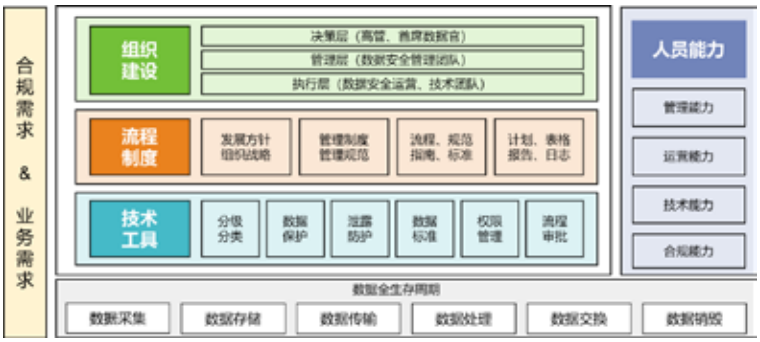
◆ 重要数据识别：

金融业机构所承载重要数据的识别和认定工作宜遵照国家及行业主管部门有关规定执行。重要数据的安全级别不宜低于本标准中确定的5级。重要数据是指我国政府、企业、个人在境内收集、产生的不涉及国家秘密，但与国家安全、经济发展以及公共利益密切相关的数据（包括原始数据和衍生数据），一旦未经授权披露、丢失、滥用、篡改或销毁，或汇聚、整合、分析后，可能造成严重后果。重要数据的性质和内容相关描述参见《标准》附录C。

金融机构的应对建议

本标准的发布可指导金融业机构明确金融数据保护对象，开展数据安全分级工作，合理分配数据保护资源和成本，从而建立完善金融数据生命周期安全框架。

绿盟科技作为国内信息安全领域的领军企业，基于多年在数据安全领域的实践经验，可协助金融机构根据《标准》的要求开展数据安全分级工作，结合绿盟科技数据安全建设体系：“一个中心，四个领域，五个阶段”完善金融机构数据全生命周期的安全防护体系。一个中心是指以数据安全防护为中心。四个领域是指的数据安全建设的四个领域：组织建设、制度流程、技术工具和人员能力。五个阶段是指的数据安全建设的五个阶段：业务梳理，分级分类，策略制定，技术管控，优化改进。



数据安全建设体系

绿盟科技数据安全建设的五个阶段，总结起来就是五个字“知”、“识”、“控”、“察”、“行”。

- ◆ 知：分析政策法规、梳理业务及人员对数据的使用规范，定义敏感数据；
- ◆ 识：根据定义好的敏感数据，利用工具对全网进行敏感数据扫描发现，对发现的数据进行数据定位、数据分类、数据分级。
- ◆ 控：根据敏感数据的级别，设定数据在全生命周期中的可用范围，利用规范和工具对数据进行细粒度的权限管控。
- ◆ 察：对数据进行监督监察，保障数据在可控范围内正常使用的同时，也对非法的数据行为进行了记录，为事后取证留下了清晰准确的日志信息。
- ◆ 行：对不断变化的数据做持续性的跟踪，提供策略优化与持续运营的服务。

借助绿盟科技在数据安全建设领域的成熟经验，可为金融机构的数据安全搭建全面可信的防御体系，有效保护数据在全生命周期过程中的安全，提升金融机构的数据安全管理和防护水平，达到合法采集、合理利用、静态可知、动态可控的数据防护目标。

数据淘金热时代下的隐私问题何去何从——探讨国内外法规下的匿名化概念

在大数据时代，数据采集和存储变得越来越容易。在政府、互联网、运营商、医疗、银行和电力等各行各业的大数据中，或多或少与个人信息有关。比如手机APP收集用户的个人注册，网页浏览息，购物和GPS位置等信息；运营商收集用户注册、电话账单、GPS以及使用流量等信息；医院会记录患者个人基本信息，以及医疗原始数据和诊断等信息。

与此同时，Open Data成为全球的大数据发展的典型趋势。数据共享、发布、外包，甚至交易等场景需求变得越来越多。欧盟在2003年实施了《公共部门信息再利用指令》；美国在2009年颁布了《开放政府指令》。我国同样实施一系列的政策和措施，比如《促进大数据发展行动纲要》（2015）、《贵州省大数据发展应用促进条例》（2016），以及今年5月份实施的《中华人民共和国政府信息公开条例》。此外，我国《网络安全法》的第十八条指出“国家鼓励开发网络数据安全保护和利用技术，促进公共数据资源开放，推动技术创新和经济社会发展”。

打破数据孤岛效应，促进数据流通，在数据流动中实现数据价值的最大化，是数据控制者/处理者的共同目标。数据蕴藏的巨大价值与隐私保护之间的矛盾日益突出，如何更好地实现两者的平衡是当今一个亟需解决的行业关键性问题，引起社会各界的强烈关注。

图1 数据利用与隐私保护两者平衡是全球的共识

匿名化作为一种解决以上困境的有效技术，在学术界首先引入深入广泛的研究，包括各类不同算法、模型以及基础理论的研究，如著名的K-匿名算法(K-anonymity)。另一方面，由于匿名化 (Anonymization) 可实现“经过处理无法识别特定个人且不能复原”，这个概念逐步被各个国家的相关立法机构所接受、所采纳。所谓匿名化，从字面理解，是匿名的处理，而“匿名”可理解是将原始数据记录代表的“自然人”实现“身份匿名”。具体来说，通过各种技术手段，比如删除标识符、泛化和加噪等操作，切断自然人与数据记录的关联。从而不仅保留所需的数据价值，同时降低了隐私泄露的风险。

图2 匿名化示例：初识概念

法规作为数据安全治理和建设的顶层指导，研究这些法规，一方面有助于更好地理解安全场景与需求，进而有利于将安全技术实际的落地与应用。近年来，全球掀起了“数据安全与隐私保护”的立法热潮，对个人信息的采集、使用和存储进行严格的法律的规范，对不合规企业进行相应等级的处罚。在其中一些法规中，匿名化处理后的数据或多或少存在一些法规的豁免权，例如可以进行共享、发布和交换。概念的定义是法规的重要基础。笔者作为技术研究人员，尝试阐述和解读欧盟、日本以及美国相关法规对匿名化概念的定义以及区别，最后笔者基于这些材料和解读发表几点不成熟的思考。国外法规和标准的翻译，笔者小心翼翼的翻译和修改，但难免存在小许的错误和偏差，欢迎各位专家指正。本文后面附上原始的英文表达。

1 国内外的匿名化相关概念定义

匿名化 (Anonymization) 相关概念 (如匿名信息 (Anonymous information)、匿名处理信息(Anonymously processed information)) 在欧盟和日本有相对成熟的

法律定义，我国近年来的法规和标准也在逐步采纳匿名化这个概念。值得关注的是，美国的相关法规中多采用一个相近的概念——去标识化 (De-identification，亦翻译为去身份化) 相关概念 (如去标识的信息，Deidentified information)。下面对相关法规进行简单的解读和分析。

1.1 欧盟

《通用数据保护条例》(General Data Protection Regulation，简称GDPR)：前言第 26 段，匿名信息 (Anonymous information) 是指与已经识别 (identified) 或可能识别 (identifiable) 的自然人不相关的信息，或者以数据主体不可或不再可识别的方式提供的信息。该段前文对“可能识别”(identifiable) 进行了解释：为了确定自然人是否可识别，应考虑合理且可能地(reasonably likely)穷举所有手段，例如由控制者或另一人单独挑选 (singling out)，以直接或间接地识别自然人。

解读：在GDPR的第四条第1款定义：“个人数据 (Personal data) 是关于一个已识别或者可能识别的自然人（即数据主体）的任何信息…”。显然，匿名信息和个人数据均建立在数据主体的识别基础。个人数据的“识别”实际包含两个层面的含义：① 已经识别 (identified)；② 可能识别 (identifiable)。已经识别 (identified) 表示这个身份信息已经确定，比如张三，11010819800101XXXX，男，出生年月1980.01.01，北京市海淀区A街道B小区，信用卡消费20W（假设该地区只有一人叫张三）；可能识别 (identified)，一般来说，一般在没有额外背景或额外身份数据库等信息的情况下定位到这个人。比如另一个处理的信息：男，出生年月1980.01.01，住在北京市海淀区A街道B小区的人，信用卡消费20W。但这不排除存在某种可能性，张三的朋友正好十分了解他的出生年月，推断出这条信息属于“张三”，进而获得了他的隐私信息。由此可见，可能识别 (identifiable)比已经识别 (identified)的判定更加苛刻，不具有已经识别的数据但仍然可能存在“可能识别”的属性。

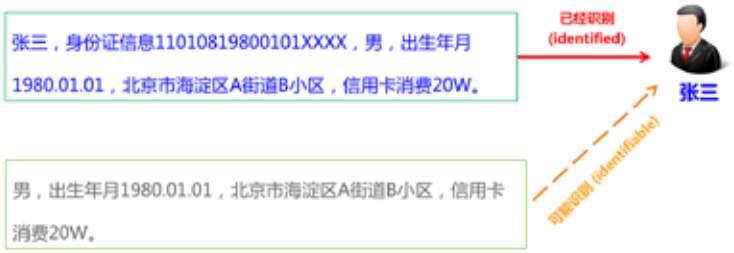


图3 GDPR语境下已经识别 (identified) 和可能识别 (identifiable) 的通俗理解

如GDPR所述，匿名信息正好相反，是一种建立消除个人数据的“已经识别”(identified)和“可能识别”(identifiable) 得到的处理结果。GDPR在该段中进行了详细解释：“为了确定自然人是否可识别，应考虑合理且可能地(reasonably likely)穷举所有手段，例如由控制者或另一人单独挑选 (singling out)，以直接或间接地识别自然人。为了确定是识别自然人的手段是否是合理且可能，应考虑所有客观因素，例如识别的成本和所需的时间，同时考虑到识别时的可用技术，技术水平和技术发展”。说明GDPR对匿名信息的判定与否，特别是对可能存在“可能识别”的一类数据，它评估和判定的手段是“合理且可能的”，而这个“合理且可能的”手段并不是几种固定的方法集合，可选的识别手段与当前“识别的成本和所需的时间，“识别时的可用技术，技术水平和技术发展”完全相关。一般来说，未来的攻击者比现在的攻击者获得的数据资源、关联技术将更多更强，这种评估是动态的。GDPR立法层面希望尽可能通过一部法律覆盖所有的个人数据安全问题，力求“大而全”。但该判定标准的描述仍然十分抽象，不够明确，为GDPR法规的实施执行留下了相应的解释空间。具体来说有以下几个问题：何为“合理地”(reasonably)？何为“可能地”(likely)？谁来做最终的判定？在下面的一个文件的观点可以找到一部分的答案。

《关于匿名化技术的意见书》：匿名处理结果的3个判断标准：(i) 是否仍有可能挑出一个人？(ii) 是否仍有可能将一个人记录关联起来？(iii) 是否可以推断有关个人的其他信息？Art.29WP 在该文件上提到：当一项提案不符合其中任意一项标准时，应对剩余的重识别风险进行彻底的评估。如果国家法律要求管理局对匿名处理程序进行评估或授权，则应向当局提供这一评估。

解读：该意见由欧盟29 条工作组 (Article 29 data protection working party，简称Art.29WP)在2014年发布。Art.29WP由每一个欧盟成员国的数据保护管理局、欧洲数据保护监督员和欧洲委员会的代表组成，代表了欧盟官方机构对数据隐私保护的权威意见。该意见书从这三个问题维度进行判断，相比GDPR的定义的“合理且可能”的限定更加严格，有追求完美的嫌疑。该意见书对假名化(Pseudonymisation)、加噪(Noise addition)、K-匿名化(K-anonymity)等多种技术进行分析，结论是这些方法处理后的数据均不符合这三个标准，多少都存在一定程度的剩余风险。例如，在K-匿名化中，每一个等价组中有K（K≥2）个实体，它不能被唯一挑选出来；但它仍然存在链接的可能性，链接成功的概率为 1/K；由于K-匿名化并没有考虑到敏感属性的分布，因此对于敏感属性相同的组，不能抵抗推断攻击。总之，它仅满足(i)，不满足(ii)和(iii)。Art.29WP并没有否定各种各样的匿名化技术实现技术，在文中评估了各种技术的优势和缺陷，指出匿名化技术在不同的场景中仍然是降低识别风险重要的措施。同时地，也明确地提出重识别风险的评估重要性。

1.2 日本

《个人信息保护法》：第二条第9款，本法中的“匿名处理信息”（Anonymously processed information）是指通过处理个人信息而产生的相关信息，它既不能根据采取以下规定的处理措施来识别到特定个人，也无法还原成个人信息。(i) 删除个人信息包含的个人描述部分等 (包括将描述部分替换为其他描述部分，或者使用具有不可恢复的方法等); (ii) 删除所述个人信息中所包含的全部标识符 (包括将标识符替换为其他描述部分，或者使用具有不可恢复的方法等)。

解读：在官网上有一个更为通俗的定义：“匿名处理信息是指，为了不能够识别特定的个人，对个人信息进行加工，并且不能复原个人信息的信息”。这个定义同样采用了“不能识别”和“不能复原”的类似描述，但和

欧盟的差别在于没有强调“识别手段”的限制条件是“合理且可能地(reasonably likely)”。那么的匿名处理才能达到“不能识别”和“不能复原”？具体怎么进行匿名处理操作呢？下面这份文件似乎提供了答案。

《个人资料保护委员会秘书处的报告：匿名处理信息》：关于匿名处理信息的处理和适当处理，由第三方组织（即个人信息保护委员会）提供最低标准，而经认可的组织等应制定个人信息保护政策和其他特定的自愿性规则，以及促进相关业务运营商遵守此类规则。期望通过这种措施来确保正确地使用个人数据，从而确保公众的安全感。

解读：该文件明确指出由个人信息保护委员会提供最低标准，同时也给出了具体的一些实践案例与指导。这与欧盟做法不同，欧盟需要自身进行评估，提交报告给相关的管理局；而日本的匿名化处理标准可直接由个人信息保护委员会提供，有一个更为统一和具体的实施标准，匿名处理信息的边界范围更加清晰，实践和操作性更强。

1.3 美国

《加州消费者隐私法案》（《California Consumer Privacy Act》，简称《CCPA》）：

“去标识”（Deidentified）指的是信息不能合理地（reasonably）识别，关联，描述，被联系在一起，或者说被链接，直接地或间接地，到特定消费者。提供商业使用的去标识的信息（满足）

(1) 已经实现了技术保护措施,并且禁止重识别（reidentification）的消费者有关的信息；

(2) 已经实现了明确地禁止信息重识别的业务流程；

(3) 已实施业务流程，以防止因疏忽而发布(reidentification)的信息；

(4) (确保) 未尝试重识别信息；

解读：美国的法规多数没有采用匿名化（Anonymization）概念，进而取代使用“去标识化”（Deidentification）相关概念，如医疗隐私相关法案HIPAA。对于CCPA去标识处理后的结果——去标识信息，与GDPR的匿名信息十分相近，但从以上定义看，存在区别：CCPA强调的“去标识信息”的识别评估手段应该是“合理的”，但没有强调是“可能的”，弱化了某些低概率的识别手段（即低概率发生的识别手段或技术）。因此，可知美国CCPA语境下的“去标识信息”更GDPR的“匿名信息”门槛更低，但这意味着前者存在的“重识别剩余风险”更高。从

上述的定义看，CCPA已经将这一类信息的使用方法和范围进行严格限定，一是通过法规限制重识别，另一个是通过技术的措施防止重识别。GDPR和CCPA给出两种完全不同的解决思路：前者处理数据门槛更高，后面的使用范围更宽；后者门槛低，后面的使用范围相对窄一些。这两者具有各自的优势所在。

1.4 中国

《网络安全法》：第四十二条 网络运营者不得泄露、篡改、毁损其收集的个人信息；未经被收集者同意，不得向他人提供个人信息。但是，经过处理无法识别特定个人且不能复原的除外。

解读：上述《网络安全法》的“经过处理无法识别特定个人且不能复原的”描述和“匿名化”(Anonymization)、“去标识化”(De-identification) 的描述，但并未明确对应两者中的哪一个。这些问题尚不明确。在我国的标准，《个人信息安全规范》中给出两者的区别：① 匿名化是通过对个人信息的技术处理，使得个人信息主体无法被识别，且处理后的信息不能被复原的过程；②去标识化 (De-identification) 通过对个人信息的技术处理，使其在不借助额外信息的情况下，无法识别个人信息主体的过程。由此可见，匿名化门槛更高一些。我国在标准中采用了两个概念，但后续法规是使用哪个概念呢？在《数据安全管理办法（征求意见稿）》，两次提出“匿名化”一词，这说明我国法规界对“匿名化”概念的在逐步接受到采纳的过程，但征求意见稿是否保留此概念，值得后续期待。何为“无法被识别”或“不能被复原”？谁来判定匿名化处理后数据的“无法被识别”或“不能被复原”？笔者也希望后续的颁发相关法规和标准能进一步解决两个基础性问题。

2 匿名化相近概念及辨析

在国内外的数据安全技术标准中，除了匿名化 (Anonymization) 和去标识化(De-identification) 概念外，我们可以看到其他两个较为相近的概念，假名化 (Pseudonymization) 和K -匿名(K-anonymity)。需要说明的是，技术标准强调是技术范畴，是处理过程，实施手段；而法规概念多数强调的处理后得到的结果或者达到的目的，如匿名信息 (Anonymous information)、去标识信息 (Deidentified information)。下面基于国内外标准的几个相关技术概念的定义进行对比和辨析。

2.1 国内标准

《个人信息安全规范》：

匿名化 (Anonymization)：通过对个人信息的技术处理，使得个人信息主体无法被识别，且处理后的信息不能被复原的过程。

去标识化 (De-identification) : 通过对个人信息的技术处理，使其在不借助额外信息的情况下，无法识别个人信息主体的过程。

解读：根据这个标准的两个定义：可看出匿名化的门槛（即处理结果的识别评估标准）比去标识化门槛更高。同时，若一个技术属于匿名化技术，那么它一定满足去标识化技术，因此去标识化包括匿名化。

《个人信息去标识化指南》

假名化 (Pseudonymization)：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。

K -匿名(K-anonymity)：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。

解读：由上看出，假名化和K -匿名一般指的是两种具体的实施技术，它们一定属于去标识化技术的范畴。而K -匿名(K-anonymity)的模型通过K个记录相同，使得攻击者无法识别该记录的“个人信息主体”，可看作实现匿名化的一种理想的手段。K -匿名一定属于匿名化技术吗？笔者认为未必，比如K -匿名中，敏感属性值相同，相当于攻击者实现了“重识别攻击”，因此不满足匿名化的目标。仅代表笔者的观点，欢迎探讨。

2.2 国外标准

《De-Identification of Personal Information》(NISTIR 8053)：

去标识化(De-identification)：表示移除一组标识数据和数据主体之间的关联的任何过程的一个通用术语。

解读：NIST定义的“去标识化(De-identification)”目标是“减少信息能够与数据主体关联的程度”而 “无法重识别”“不可复原”（匿名化目标）。在该标准中解释到：使用“去标识化 (De-identification)”一词，是指有时可以实现重识别，有时则不能”。总之，可得出结论：中国和美国的标准中，去标识化的门槛比匿名化的更低，因此范围更广，它们存在包含关系。

《Privacy enhancing data de-identification terminology and classification of techniques》(ISO/IEC 20889)：

假名化 (Pseudonymization) 是去识别化技术的一种，它将数据主体的标识符(或一组标识符)替换为假名，以隐藏该数据主体的身份。

解读：假名化是实现去标识化的一种方法。其将数据主体的标识符(或一组标识符)替换为假名，假名可由随机的替换表、哈希函数、加密算法实现获取。

《Privacy enhancing data de-identification terminology and classification of techniques》(ISO/IEC 20889)：

K-匿名 (K-anonymity) 是指一种隐私度量模型，它确保数据集中的每个标识符都有一个对应的等价类，且等价类中至少包含K条记录。

解读：K-匿名同样是实现去标识化的一种方法。它对准标识化进行泛化和处理，使得攻击者无法从准标识进行链接攻击，唯一识别到某个数据主体相关的记录，从而保护了敏感属性。一定程度地说，K-匿名通过数学模型的限制，能相比其他去标识化方法（假名化、加噪）等更能逼近法律的匿名化目标，但它也只是实现匿名化的一个技术手段。

2.3 范畴关系

通过以上的概念比较，笔者认为它们之间的关系可用下图表示。需要进一步解释的是，由于K-匿名模型，比如在某个等价组中敏感属性完全相同，仍然存在隐私属性泄露风险，即可看作实现了“重识别攻击”。因此笔者认为，它不一定满足法规和标准中对匿名化概念的定义。抛砖引玉，欢迎探讨。

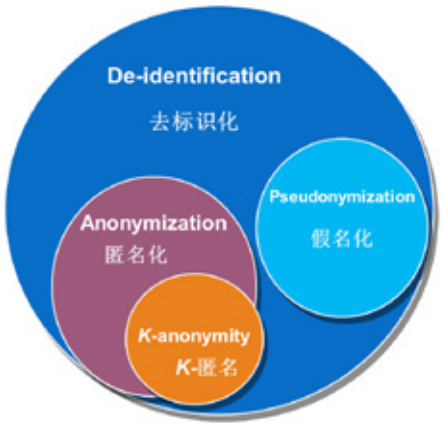


图4 几个概念之间的关系

3 小结

通过对国外法规标准的研究，可得到以下一些重要结论：

(1) 欧盟 (GDPR) 和日本 (《个人信息保护法》)在法规中多采用匿名化 (Anonymization) 相关概念；

(2) 美国的法规（如HIPAA,CCPA）采用去标识化 (De-identification) 的相关概念；

(3) 美国的CCPA对去标识信息 (de-identified information)定义比欧盟的GDPR对匿名信息 (Anonymous information) 的定义门槛更低，但CCPA对此类数据做了更多的限制，通过法规和技术措施防止重识别；GDPR语境下的匿名信息不是个人信息，不受GDPR的重重管制；

(4) 欧盟的GDPR的对匿名信息的判定“合理且可能”的识别手段，《关于匿名化技术的意见书》可看出需要企业向相关的管理局提供评估报告。日本的做法与欧盟不同，在《个人信息保护法》及标准明确指出由个人信息保护委员会提供最低标准，标准更加统一且具体。

进一步地，作为一个技术出身的数据安全从业者，十分关注国内的数据安全的法规环境、以及技术动态。通过对比研究，从三个层面发表几点不成熟的意见，抛转引玉，欢迎各位专家探讨：

(1) 在立法层面：我国在《网络安全法》中并没有明确提出匿名化和去标识化概念，在《数据安全管理办法（征求意见稿）》首次引入“匿名化”概念；但没有给出法规的定义。希望后续的法规《个人信息保护法》、《数据安全法》等能明确匿名化或去标识化概念的具体定义。此外，结合我国大数据以及数据安全的发展现状，借鉴和吸收欧盟和美国对匿名/去标识数据的两种管理方式，在数据利用和数据安全进行平衡，完善匿名化相关制度的设计。

(2) 在技术层面：两种基础技术。① 开展隐私风险评估技术研究。目前国外已有一些数据集的重识别风险评估研究，我国目前几乎处于空白。目前有多种技术手段，包括数据脱敏、匿名化、假名化和差分隐私等等，去实现法规中的“匿名化”或“去标识化”。然而，那种方法降低的隐私风险更低呢？目前业界缺乏一

个统一的评判标准；② 实用可控的匿名化技术研究。目前的技术手段并不能很好应付各种各样的数据开放、共享和发布场景。比如高维数据集，关联关系，效率问题，自适应场景问题，最优平衡等等，均是推动实用化进程中亟需解决的关键性问题。

(3) 在应用层面：匿名数据/去标识数据的处理和使用，不仅要通过技术防护，也要加强管理措施。如 ① 企业内部的数据共享：个人隐私信息使用假名化技术处理，保留了数据更多的特性，同时对用户的身份信息数据库进行严格访问权限管理，同时严格使用者的使用频次和时间，从各个因素上控制和降低隐私泄露风险；② 企业间的数据共享：数据控制方将包含个人信息的数据外包给第三方，根据需求对数据进行严格的匿名处理手段，同时通过双方签订使用协议，限制通过使用高级技术的实现对个人信息的重识别；③ 数据完全对开放：数据控制方将数据对外发布，由于潜在的风险很大，因此需加强对匿名化处理数据，加强评估力度，通过算法评估指标以及专家抽查等方式反复进行评估，确保风险在控制范围内。

参考文献

[1]《中华人民共和国网络安全法》http://xxzx.mca.gov.cn/article/wlaqf2017/wjjd/201705/20170500891068.shtml

[2]《数据安全管理办法》（征求意见稿）http://www.moj.gov.cn/news/content/2019-05/28/zlk_235861.html

[3]《General Data Protection Regulation》, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv: OJ.L_.2016.119.01.0001.01.ENG&toc=OJ:L:2016:119:TOC

[4] Article 29 Data Protection Working Party: Opinion 05/2014 on Anonymisation Techniques

[5]《個人情報保護法》https://www.ppc.go.jp/files/pdf/290530_personal_law.pdf

[6] 匿名加工情報制度について https://www.ppc.go.jp/personalinfo/tokumeikakouInfo/

[7] 個人情報保護委員会事務局レポート：匿名加工情報，2017,2

[8]《California Consumer Privacy Act》, https://cal-privacy.com/

[9] 王融. 数据匿名化的法律规制[J]. 信息通信技术, 2016, 10(4): 38-44.

[10] 韩旭至. 大数据时代下匿名信息的法律规制[J]. 大连理工大学学报: 社会科学版, 2018, 39(4): 64-75.

附录—原始定义

欧盟 GDPR 前言 26 段对匿名化的相关描述：

“The principles of data protection should apply to any information concerning an identified or identifiable natural person. Personal data which have undergone pseudonymisation, which could be attributed to a natural person by the use of additional information should be considered to be information on an identifiable natural person. To determine whether a natural person is identifiable, account should be taken of all the means reasonably likely to be used, such as singling out, either by the controller or by another person to identify the natural person directly or indirectly. To ascertain whether means are reasonably likely to be used to identify the natural person, account should be taken of all objective factors, such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of the processing and technological developments. The principles of data protection should therefore not apply to anonymous information, namely information which does not relate to an identified or identifiable natural person or to personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable. This Regulation does not therefore concern the processing of such anonymous information, including for statistical or research purposes”.

日本匿名报告的英文版 Report by the Personal Information Protection Commission Secretariat: Anonymously Processed Information

https://www.ppc.go.jp/files/pdf/The_PPC_Secretariat_Report_on_Anonymously_Processed_Information.pdf Article 2 (9)对匿名信息的定义：

"Anonymously processed information" in this Act means information relating to an individual that can be produced from processing personal information so as neither to be able to identify a specific individual by taking action prescribed in each following item in accordance with the divisions of personal information set forth in each said item nor to be able to restore the personal information.

(i) personal information falling under paragraph (1), item (i); Deleting a part of descriptions etc. contained in the said personal information (including replacing the said part of descriptions etc. with other descriptions etc. using a method with no regularity that can restore the said part of descriptions etc.)

(ii) personal information falling under paragraph (1), item (ii); Deleting all individual identification codes contained in the said personal information (including replacing the said individual identification codes with other descriptions etc. using a method with no regularity that can restore the said personal identification codes)

美国 CCPA 对去标识信息的定义

“Deidentified” means information that cannot reasonably identify, relate to, describe, be capable of being associated with, or be linked, directly or indirectly, to a particular consumer, provided that a business that uses deidentified information:

(1) Has implemented technical safeguards that prohibit reidentification of the consumer to whom the information may pertain.

(2) Has implemented business processes that specifically prohibit reidentification of the information.

(3) Has implemented business processes to prevent inadvertent release of deidentified information.

(4) Makes no attempt to reidentify the information.

NSFOCUS

技术实践

个人金融信息保护视角下的脱敏效果评估研究与实践

绿盟科技：施岭、陈磊

1 引言

2月13日中国人民银行发布《个人金融信息保护技术规范》的行业标准，指导各相关机构规范处理个人金融信息，最大程度保障个人金融信息主体合法权益，维护金融市场稳定。

《个人金融信息保护技术规范》中对个人金融信息做了定义，指包括账户信息、鉴别信息、金融交易信息、个人身份信息、财产信息、借贷信息和其他反映特定个人金融信息主体某些情况的信息。本规范中根据信息遭到未经授权的查看和未经授权的变更后所产生的影响和危害，将个人金融信息按敏感程度从高到低分为C3、C2、C1三个类别。同时基于个人金融信息在进行收集、传输、存储、使用、删除、销毁等生命周期各环节的处理过程提出了安全防护要求，并从安全技术和安全管理两个方面，对个人金融信息保护提出了规范性要求。

2 基于个人金融信息使用安全的要求

在《个人金融信息保护技术规范》规范中提到了“匿名化”和“去标识化”两种技术，并对这两种技术效果做了解释：

- **匿名化：**指通过对个人金融信息的技术处理，使得个人金融信息主体无法被识别，且处理后的信息不能被复原的过程。注：个人金融信息经匿名化处理所得的信息不属于个人金融信息。
- **去标识化：**指通过对个人金融信息的技术处理，使其在不借助额外信息的情况下，无法识别个人金融信息主体的过程。注：去标识化仍建立在个体基础之上，保留了个体颗粒度，采用假名、加密、加盐的哈希函数等技术手段替代对个人金融信息的标识。

从以上两个定义中可以看出“匿名化”与“去标识化”的区别指信息处理后是否能被复原，信息处理后是否属于个人金融信息。

我们从另一个角度来看这两种技术，“匿名化”和“去标识化”又可以简单理解为脱敏的两种技术，其尤其针对个人隐私保护起到重要的作用。那么在数据使用安全的环节中，提到了多种场景，包括：信息展示、共享与转让、公开披露、委托处理、加工处理、汇聚融合、开发测试，这些场景最好的安全防护措施就是利用权限对信息内容进行处理后再使用，那么信息内容处理中最优的方法就是数据脱敏，而脱敏效果的优劣程度，就成为了信息使用准确性、有效性的重要考量因素，因此在信息处理后应进行科学的脱敏效果评估。

本文主要阐述个人金融信息在生命周期各环节信息处理后的安全性评估，个人金融信息通过处理后的安全性评估可以帮助金融机构更好的实现个人金融信息在后续使用中的安全，还可以验证信息处理后的质量。

3 数据脱敏效果评估从理论到实践

本章节下面将首先介绍重标识风险评估相关的场景与理论，重点介绍国外主流的评估方法与评估指标。为了阐述评估方法的应用，将在个人金融脱敏数据集进行实践与应用（需要说明的是，数据脱敏效果也称为重标识风险评估，前者较多使用在工业界，后者较多使用在学术界，本文将根据语境混用这两个词）。

3.1 重标识风险评估场景与理论

3.1.1 脱敏数据集的重标识攻击

Open Data成为全球的大数据发展的典型趋势。数据共享、发布、外包，甚至交易等场景需求变得越来越多。为了降低数据敏感性，一般会经过脱敏处理，再进行发布和共享。然而，在大数据时代，一些以前被认为数据脱敏技术，比如各种去标识化方法（泛化、屏蔽、假名、删除标识符等）仍然存在重新识别身份攻击风险，这种攻击通常称为“重标识攻击”。

历史上一个经典的重标识攻击与隐私窃取案例：1996年，美国马萨诸塞州发布了医疗患者信息数据库（DB1），去掉患者的姓名和地址信息，仅保留患者的{ZIP, Birthday, Sex, Diagnosis,…}信息。另外有另一个可获得的数据库（DB2），是州选民的登记表，包括选民的{ZIP, Birthday, Sex, Name, Address,…}详细个人信息。攻击者将这两个数据库的同属性段{ ZIP, Birthday, Sex}进行链接操作，可以恢复出大部分选民的医疗健康信息，从而导致一起严重的选民医疗隐私数据泄露事故。

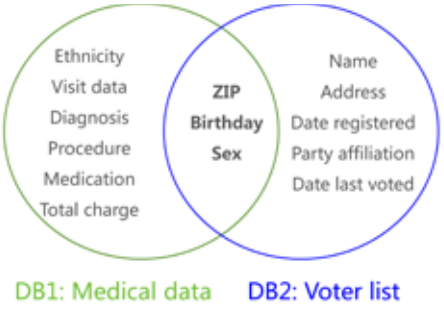


图1 链接攻击（Linking attack）示意图

3.1.2 重标识风险评估理论

对于脱敏数据集的重标识风险评估问题，最为简单方法是使用脱敏数据集

的唯一性(uniqueness)指标进行度量。若某条记录的单个属性或者多个属性组合的值在脱敏数据表中是唯一的，那么很可能被攻击者唯一关联出个人信息主体的身份。据权威机构统计，在美国，使用邮编、性别、出生日期信息，有81%的概率可以唯一识别出对应的美国公民。对于大规模数据集的唯一性度量问题， Google学者Chia等人在2019年的IEEE Symposium on Security and Privacy会议上提出了适应大规模数据集（PB级别）场景的重标识评估算法。在2019年的Nature communications期刊上，Rocherd等人对于发布的数据集重标识评估问题，提出重标识评估的近似模型模型，成功刻画了去标识化/假名化数据集被重标识的成功率，证明即使是不完整的数据集 (比如社会人口调查采样的数据集)，某些个人数据属性的组合仍然存在高的重标识风险。

在重标识风险评估领域，多个学者进行一系列的指标与评估算法的研究，其中较为著名的要数加拿大学者El Emam，他进行了深入的研究，他提出了三种常见的攻击场景以及一系列的评估指标，并将将研究成果进行了转化，创立一家科技公司——Privacy Analytics，主要面向医疗隐私数据，并对去标识化结果进行风险评估与检测，帮助数据处理企业合规美国医疗HIPAA法案，同时获得最大的数据价值，比如将评估合规的脱敏医疗数据出售给保险、药企和科研结构等第三方。

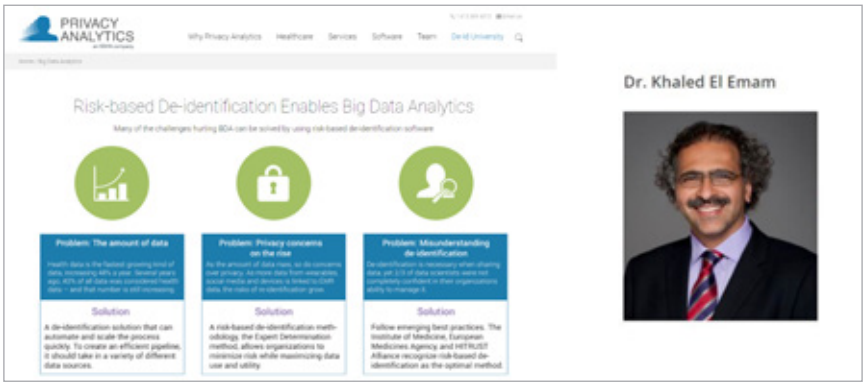


图2 Privacy Analytics 网站主页以及创立者Khaled El Emam

El Emam根据攻击者的目的和攻击能力定义了三类常见的隐私攻击场景——并形象化地被称为检察官攻击 (Prosecutor attack)、记者攻击 (Journalist attack)和营销者攻击 (Marketer attack)。其中检察官攻击具有背景知识，了解攻击目标一定在公开数据集中，比如攻击者了解自己的朋友在发布展示的数据集中，他的目的是挑选出他的朋友，并获得敏感属性信息（比如财产消费、医疗健康等信息）；记者攻击场景是达到曝光的目的，他需要尽量寻找公开数据库（比如选举身份登记表），进行匹配关联，进行多次向媒体炫耀，证明重新识别个体，使得涉及企业名誉扫地和难堪；营销者攻击目的是营销，即企业对自己的用户进行多维度的关联与画像，只需保持较高识别概率的匹配关联即可，无需证明是唯一识别出脱敏数据集对应的数据主体。

攻击场景	描述
检察官攻击	攻击者知道某个特定人员在公开的数据集中发生的重标识攻击，他发起的攻击是指向特定目标的，例如某个人了解他的同学是受访对象，他去查找他的同学属于公开的去标识化数据集的哪一行记录
记者攻击	在此场景中，攻击者一般来说拥有一个大的身份数据库，但他并不知道数据库的人员是否在公开的数据集中，他通过多次炫耀式攻击证明某人可以被重新识别。在这种情况下，攻击者的目标常常是使得公开数据库的组织感到难堪或者名誉扫地
营销者攻击	类似记者攻击场景，但攻击者的目标是使得公开数据库和身份数据库进行关联下实现的重标识攻击。尽量还原出公开数据库的身份，实现精确对身份数据库的人进行其他维度的刻画，但不要求证明重标识结果的正确性，只需保证较高的重标识概率

图3 三种重标识攻击场景

在检察官攻击、记者攻击和营销者攻击三类攻击场景下，El Emam学者等根据脱敏数据集的最高、平均重标识风险的概率、高风险记录占有比例等刻画需求，设计了8类评估指标，具体如图4所示。这些指标可有效地评估结构化的脱敏数据集的隐私风险残余情况。这些指标的数值范围均为[0,1]，1表示最高风险，0表示几乎无风险。

风险类型	评估指标	指标意义	符号含义
检察官攻击风险	$_{p}R_{a}=\frac{1}{n}\sum_{j\in J}f_j\times I\left(\frac{1}{f_j}>\tau\right)$ $_{p}R_{b}=\frac{1}{\min_{j\in J}(f_j)}$ $_{p}R_{c}=\frac{\lvert J\rvert}{n}$	$_{p}R_{a}$ 刻画重识别概率大于 τ 的数据集记录占总体的比例； $_{p}R_{b}$ 刻画数据集所有记录中最大的重识别概率； $_{p}R_{c}$ 刻画平均重识别概率。	①n—数据集记录的数量； ②J—数据集的等价组的集合； ③$\lvert J\rvert$—数据集的等价组数量； ④f_j—数据集等价组为 $j\in J$ 的数量； ⑤τ—阈值； ⑥$I(\cdot)$—当输入为真，输出为1，否则为0； ⑦N—身份数据集记录（可访问或拥有的）的数量； ⑧F_j—身份数据集（可访问或拥有的）等价组为 $j\in J$ 的数量。
记者攻击风险	$_{j}R_{a}=\frac{1}{n}\sum_{j\in J}f_j\times I\left(\frac{1}{F_j}>\tau\right)$ $_{j}R_{b}=\frac{1}{\min_{j\in J}(F_j)}$ $_{j}R_{c}=\max\left(\frac{\lvert J\rvert}{\sum_{j\in J}F_j},\frac{1}{n}\sum_{j\in J}\frac{f_j}{F_j}\right)$	$_{j}R_{a}$ 刻画重识别概率大于 τ 的数据集记录占总体的比例； $_{j}R_{b}$ 刻画数据集所有记录中最大的重识别概率； $_{j}R_{c}$ 刻画平均重识别概率。	
营销者攻击风险	$_{M}R_{1}=\frac{\lvert J\rvert}{N}$ $_{M}R_{2}=\frac{1}{n}\sum_{j\in J}\frac{f_j}{F_j}$	$_{M}R_{1}$，$_{M}R_{2}$ 分别刻画在情况1和2下的平均重识别概率； 情况1：身份数据集和发布数据集的个人信息主体完全相同； 情况2：发布数据集是身份数据集的个人信息主体的一部分。	

图4 三种重标识攻击场景

3.2 在个人金融数据集的评估实践

3.2.1 脱敏数据集介绍

为了检验重标识风险评估指标在真实场景的有效性，我们收集了真实的23110条身份证号+手机号的记录信息，为了使得数据更具有隐私攻击价值（攻击

者的攻击意愿很强），我们随机生成了一列银行卡账户余额（40-120万元），即银行的个人金融数据集场景。为了避免不必要的隐私泄露，下面展示的三行身份证号+手机号数据均是伪造生成，不具有真实意义。

为了评估常见脱敏方法与规则的脱敏效果，下面选择了8种脱敏规则并且得到相应的8种不同的脱敏数据集。可以看出，脱敏规则Rule 1-5脱敏强度逐步加强。Rule 2-c、3-c和5-c分别是Rule 2、3和5对照组（Control Group），即脱敏/屏蔽的数字个数相同，但位置略有不同，具体规则和策略参考图5-7的脱敏表示意图。

3.2.2 重标识风险评估实践

下面我们对8种脱敏数据集在检察官攻击、记者攻击和营销者攻击三种场景下分别进行评估，然后对风险评估的结果进行定性的分析和讨论。

在检察官攻击场景中，意味着攻击者（比如可以获取脱敏数据的银行内部员工）了解你在的城市某一家银行办了一张银行卡，且他掌握了你的身份证号和手机号信息，他去查询公开的脱敏数据集，查看那一条记录属于你，从而可以窃取隐私数据——账户余额。下面图5展示不同的脱敏强度下的检察官攻击风险。

脱敏规则	脱敏数据集	pRa (高风险占的比例)	pRb (记录最高风险)	pRc (平均风险)
Rule1		0.9998	1.0000	0.9999
Rule2		0.9977	1.0000	0.9988
Rule3		0.6235	1.0000	0.6921
Rule4		0.1259	1.0000	0.1827
Rule5		0.0064	1.0000	0.0233

(a)

脱敏规则	脱敏数据集	pRa (高风险占的比例)	pRb (记录最高风险)	pRc (平均风险)
Rule2-c		0.9367	1.0000	0.9628
Rule3-c		0.4295	1.0000	0.5657
Rule5-c		0.0247	1.0000	0.0603

(b)

图5 不同脱敏数据集的检察官攻击风险：(a) 数据集的脱敏强度逐步较强；(b) 图(a)的部分对照组

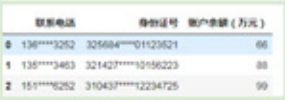
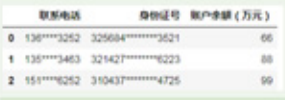
由图5的评估结果进行分析，可得到一下一些结论：

- 脱敏强度越强，重标识风险越低。在实际应用中，应根据隐私保护的需求，进行风险评估，使得隐私风险在预期范围内，保持数据的可用性。
- 仅对手机号中间4位，以及身份证的出生日期进行脱敏/屏蔽处理（Rule1-2），它们脱敏后仍然具有较高的重标识风险，检察官攻击发

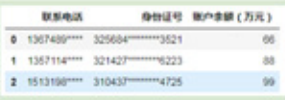
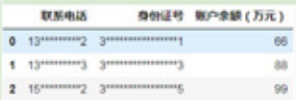
生的风险较高等级，平均风险、高风险占的比例均接近于1.0。因此，在数据公开发布场合建议不予采用。

- 仅对手机号最后8位，以及身份证的出生日期进行脱敏/屏蔽处理（脱敏规则Rule3），能有效降低隐私风险，平均重标识风险能降低至0.6921。
- 仅保留手机号的前3位，与身份证号的前2位，重标识风险能降到更低，平均风险为0.0233。在数据展示和验证场景，比如App界面展示，建议采用最小够用原则，屏蔽更多的数字码，用户能了解是自己的手机号和身份号即可。
- Rule2-c相对Rule2，平均识别风险稍有下降，这是由于屏蔽手机号最后4位比中间4位更有效，最后4位数字更有“个性”（用户号码）。在实际展示场景中，建议尽量屏蔽最后4位。

在记者攻击场景中，一个脱敏数据集已经对外公开发布，且攻击者可以公开访问或者掌握了身份数据库（比如黑客掌握一系列黑灰产数据库），他根据自己拥有的数据库去查询和关联记录对应的身份主体，查询不同身份主体拥有多少账户余额（窃取隐私数据）。这种情况一般在实际中尽可能扫描黑市的大型泄露数据库、或公开的数据库，进行比对和评估。本文为了简化，假设黑客掌握的数据库为公开脱敏数据集的10%（随机抽取10%）。在此场景假设条件下，分别给出不同脱敏数据集的评估结果，如图6所示。

脱敏规则	脱敏数据集	jRa (高风险占的比例)	jRb (记录最高风险)	jRc (平均风险)
Rule1		0.0999	1.0000	1.0000
Rule2		0.0998	1.0000	1.0000
Rule3		0.0733	1.0000	0.8260
Rule4		0.0203	1.0000	0.3582
Rule5		0.0014	1.0000	0.0930

(a)

脱敏规则	脱敏数据集	jRa (高风险占的比例)	jRb (记录最高风险)	jRc (平均风险)
Rule2-c		0.0989	1.0000	0.9917
Rule3-c		0.0640	1.0000	0.8424
Rule5-c		0.0057	1.0000	0.2163

(b)

图6 不同脱敏数据集的记者攻击风险：(a) 数据集的脱敏强度逐步较强；(b) 图(a)的部分对照组

由图6(a-b) 的评估结果可得上一小节类似的结论，即脱敏强度越强，重标识风险越低，在实际应用中，尽量在可用的情况下，屏蔽更多的数字码。

在个人信息展示场景，尽量屏蔽手机号的最后4位和身份证最后4位。

在营销者攻击场景中，类似记者攻击场景需假设营销者（攻击者）拥有的身份的数据集，但该场景的目标是为了营销和用户画像，关注关联的准确率，而不关注那些记录被正确识别，因此只评判平均重标识风险（不检测高风险记录的比例）。为了演示方便，假设营销者掌握的数据集是大的身份数据集，脱敏数据集为它表示身份主体的一部分（子集），是其中的营销者掌握的数据集记录的20%（随机抽取20%）。在此场景假设条件下，分别给出不同脱敏数据集的评估结果，如图10所示。

脱敏规则	脱敏数据集	mR2 (平均风险)
Rule1	联系电话 身份证号 账户余额（万元） 0 136****3252 325684****01123521 66 1 135****3463 321427****10156223 88 2 151****6252 310437****12234725 99	0.9997
Rule2	联系电话 身份证号 账户余额（万元） 0 136****3252 325684********3521 66 1 135****3463 321427********6223 88 2 151****6252 310437********4725 99	0.9990
Rule3	联系电话 身份证号 账户余额（万元） 0 136******** 325684********3521 66 1 135******** 321427********6223 88 2 151******** 310437********4725 99	0.6921
Rule4	联系电话 身份证号 账户余额（万元） 0 136******** 325684********** 66 1 135******** 321427********** 88 2 151******** 310437********** 99	0.1843
Rule5	联系电话 身份证号 账户余额（万元） 0 136******** 32********** 66 1 135******** 32********** 88 2 151******** 31********** 99	0.0227

(a)

脱敏规则	脱敏数据集	mR2 (平均风险)
Rule2-c	联系电话 身份证号 账户余额（万元） 0 1367489**** 325684********3521 66 1 1357114**** 321427********6223 88 2 1513198**** 310437********4725 99	0.9668
Rule3-c	联系电话 身份证号 账户余额（万元） 0 136******** 3256841989******** 66 1 135******** 3214271989******** 88 2 151******** 3104371989******** 99	0.5655
Rule5-c	联系电话 身份证号 账户余额（万元） 0 13*********2 3**********1 66 1 13*********3 3**********3 88 2 15*********2 3**********5 99	0.0609

(b)

图7 不同脱敏数据集的营销者攻击风险：(a) 数据集的脱敏强度逐步较强；(b) 图(a)的部分对照组

由图7 (a-b) 的评估结果可得上一小节类似的结论，因此不再赘述。综上三种攻击场景的风险结果可看出，风险趋势是类似的，即与数据集的概率分布密切相关。

总结

银行的个人金融信息，由于涉及用户交易、消费和财产等具有高度的敏感性。为了促进个人金融信息数据集的挖掘、利用与流通，数据脱敏（包括去标识化、匿名化）通常被作为一种重要的隐私保护手段，然而采取不同的脱敏方法以及脱敏程度，可能并没有完全彻底消除重标识的残余风险（Residual Risk）。大数据时代，攻击者/黑客通过网络入侵攻击，或者黑灰产收集和掌握的用户数据越来越多，同时去匿名、重标识、“人肉搜索”等攻击技术不断更新与迭代，重标识攻击的威胁发生频率将越来越普遍，尤其是对于高价值密度的数据集，包括个人金融数据集、医疗数据集。因此，数据脱敏技术手段处理后的数据集不应被看作是一劳永逸的，数据控制者/数据处理者应该经常性的再评估匿名化后的数据是否存在新的风险。实际上，《个人金融信息保护技术规范》、《个人信息安全规范》以及《个人信息去标识化指南》等一系列标准也明确指出实用标识化和匿名化等手段脱敏数据后，应进行评估和定期重新评估数据脱敏的效果（重标识风险）。基于此，本文阐述了数据脱敏的理论以及实践应用，以期给出应用在个人金融脱敏数据集的一般化评估方法与步骤。

参考文献

1. 绿盟科技，数据安全白皮书2.0. https://www.nsfocus.com.cn/html/2019/350_0927/49.html.

2. 2019 网络安全观察. http://blog.nsfocus.net/wp-content/uploads/2020/01/2019-Cybersecurity-Insights.pdf.

3. 陈磊，隐私合规视角下的数据安全建设与思考，保密科学与技术

4. JR/T 0171—2020， 个人金融信息保护技术规范

5. GB/T 35273-2020，信息安全技术 个人信息安全规范

6. GB/T 37964-2019，信息安全技术 个人信息去标识化指南

7. Rocher L, Hendrickx J M, De Montjoye Y A. Estimating the success of re-identifications in incomplete datasets using generative models. Nature communications, 2019, 10(1): 1-9.

8. El Emam K. Guide to the de-identification of personal health information. Auerbach Publications, 2013.

9. Article 29 Data Protection Working Party, Opinion 05/2014 on Anonymisation Techniques, 2014.

大数据下的隐私攻防：数据脱敏后的隐私攻击与风险评估

在大数据时代下，数据被大量的收集、处理、加工、提取和挖掘，其中一类与个人相关的隐私数据，比如个人基本信息、财产信息、搜索信息、社交信息、网购行为信息等具有广泛的挖掘价值，可用于个人的精确画像、定向推广等2C业务开展。伴随大数据的采集、存储和应用的增长，个人隐私面临的威胁与挑战日趋严峻——既有来自内部员工的非法查询与拷贝，也有外部黑客攻击导致的重大数据泄露。

为了平衡数据利用与隐私保护的双重需求，数据脱敏成为当前绝大数企业在数据安全治理与建设过程中的必选技术与措施，在数据分析、产品测试、培训等场景中有广泛应用。然而，数据脱敏是否真的是解决大数据时代下隐私问题与挑战的“一剂特效良药”（一劳永逸，药到病除）？或者说，脱敏后的数据是否真的达到了绝对的隐私安全（任何攻击者均无法攻破脱敏数据集）？如果不是，那么不同是脱敏方法、脱敏产品/系统的真实效果如何？如何比较它们的脱敏效果（隐私保护能力）？这引出一个重要的研究课题——数据脱敏的效果评估问题，即脱敏数据集的残余隐私风险如何定量刻画的问题。本文是大数据下的隐私攻防系列文章的第一篇《数据脱敏后的隐私攻击与风险评估》，笔者尝试将国外相关的评估技术与理论进行梳理和分析；后续第二篇《身份证号+手机号如何脱敏才有效？》，将在真实的身份证号+手机号脱敏数据集上实践与应用。通过系列文章希望达到了解风险、分析风险，刻画风险并最后控制风险的目的。抛砖引玉，以期进一步推动数据脱敏系统/产品组合的进一步完善与闭环。

1 背景

据国外情报机构Risk Based Security（RBS）2019年Q3季度的报告统计，数据泄露近年来呈爆发和递增的趋势，如图1所示：2019年公开披露的泄露事件数量达到5183次！累计泄露数据记录规模达到79亿条！

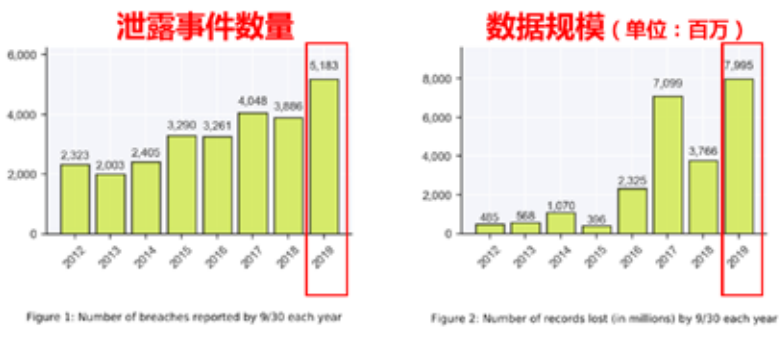


图1 全球数据泄露事件数量和规模概览（图来源于RBS报告[1]）

据《2019 网络安全观察报告》[2]对2019年具有代表性的数据泄露事件进行收集与分析，结论指出大规模的数据泄露事件泄露的数据类型基本为与个人相关的数据，涉及互联网、金融、医疗、航空、教育等2C服务的行业领域。

图2 2019年国内外披露的数据泄露事件[2]

数据泄露影响与危害，一方面给受害企业带来名誉和形象的损失；另一方面将可能面临法律的处罚。近年来，为了保障公民的个人信息与隐私安全，全球各个国家掀起了立法热潮，包括美国的GDPR、美国加州的CCPA、巴西的LGPD。最为严格的隐私法规——欧盟GDPR，要求数据控制者和处理者（企业）必须履行保护用户数据的义务，若违反将面临最高2000万欧元或4%的全球营业总额的高昂罚款。

在法规监管的背景下，企业不得不重新审视个人信息安全与隐私保护在数据安全建设的重要性和紧急性。最容易想到的隐私保护手段是对数据进行加密，然而隐私保护达到了100%，但数据的可用性几乎为0，严重阻碍了数据流通与价值挖掘。个人信息利用与个人信息保护（隐私保护），如同天平的两侧，如何平衡两者关系，尤为重要。其中，数据脱敏通过对敏感数据进行变形和失真——降低敏感度，成为当前企业在数据安全治理与建设过程中的比较青睐的技术与产品，在数据分析、产品测试、培训等企业场景有高频的应用。

2 数据脱敏与隐私攻击

2.1 数据脱敏

数据脱敏 (Data Masking)，也称为数据漂白。由于处理高效且应用灵活等特点，是目前工业界广泛处理敏感类数据采用的一种技术，在互联网、金融、运营

商、企业等有广泛应用。脱敏后的数据是否可恢复可分为可逆脱敏和不可逆脱敏。

可逆脱敏包括置换、唯一替换和保留格式加密（Format-Preserving Encryption，FPE）等。置换是指通过密钥生成的伪随机序列，对敏感数据表同一列属性值进行重排，使得攻击者无法找到“自然人”对应的准确属性信息；唯一替换，企业通过建立一些敏感词的映射表替换为其他非敏感数据，通过反向映射表可将脱敏数据恢复为原始数据；FPE是一种特殊的加密方式，其输出的密文格式仍然与明文相同，即加密过程中考虑格式及分段约束，这与传统的分组密码是不同的。比如中国联通手机号15266661234，通过FPE加密可以实现仍然输出的是联通手机号15173459527。为了规范FPE技术实施，美国NIST发布了FF1标准算法，可用于保险号、银行卡号、社保卡号等数字标识符的加密，同时适应于支付卡行业数据安全标准 (PCI DSS, Payment Card Industry Data Security Standard)。

不可逆脱敏策略有多样，可以看作失真和变形一系列工程化方法的集合，包括取整、量化、泛化、屏蔽、截断、散列和加噪等，如下表所示。具体使用哪种脱敏方法，需要根据业务场景，如数据的使用目的、以及脱敏级别等需求去选择和调整。

方法/策略	描述	示例
取整	数值或日期数据的取整	13:25:15 → 13:00:00
量化	通过量化间距调整数据失真程度	27 → 30
泛化	对数据进行抽象	海淀区→ 北京市
屏蔽	屏蔽部分数据，如电话、身份证号码	152****1234
截断	数据尾部截断	010-88886666 → 010
散列	将输入映射为固定长度的字符串	8 → a17d 28 → 1c4a
加噪	将数据加入一些噪声	12.4 → 12.9

图3 常见脱敏方法/策略

按照使用场景，可将脱敏分为静态脱敏(Static Data Masking, SDM)、动态脱敏(Dynamic Data Masking, DDM)，本文介绍是前者。静态脱敏一般用于非生产环境中（测试、统计分析等），当敏感数据从生产环境转移到非生产环境时，这些原始数据需要进行统一的脱敏处理，然后可以直接使用这些脱敏数据；动态脱敏一般用于生产环境中，在访问敏感数据当时进行脱敏，根据访问需求和用户权限进行“更小颗粒度”的管控和脱敏。一般来说，动态脱敏实现更为复杂。脱敏在多个安全公司已经实现了应用，IBM，Informatica公司是比较著名的代表。

2.2 隐私攻击

Open Data成为全球的大数据发展的典型趋势。数据共享、发布、外包，甚至交易等场景需求变得越来越多。为了降低数据敏感性，一般会经过脱敏处理，再进行发布共享。然而，在大数据时代，一些以前被认为有效的隐私保护与匿名化技术正“捉襟见肘”，仍然可能存在各种隐私攻击风险。

数据开放和共享范围不同，潜在的攻击者不同，他们的背景知识、攻击能力和攻击动机也是不同的。通常，脱敏数据集仍然存在链接攻击、重识别攻击、背景知识攻击、同质性攻击和近似性等多种隐私攻击挑战。即攻击者通过一种攻击或者攻击组合从脱敏数据集中分析、推断和获得一些相关的隐私和敏感信息。

比如图4中的背景知识攻击，假设小王了解自己的朋友“张三”和“李四”在某银行办理了信用卡，他知道朋友的一些基本信息，比如年龄和身高、毕业学校，通过背景知识和网站公开的脱敏表进行比对，重新识别出两

位朋友的在哪一行记录，并获取朋友的隐私信息——信用卡消费记录（注图4的脱敏数据集（4行记录）只是一个示意图，实际发布场景可能有100条记录以上，小王同样根据数据集的唯一特性识别出他的朋友）。

图4 背景知识攻击：小王识别出两位朋友的记录，获得隐私信息

攻击场景和攻击者的能力假设条件必须是合情合理的，即在真实场景是可能发生的。下面给出两个披露的在脱敏数据集实施隐私攻击的经典案例：

案例1：美国医疗隐私泄露（两个数据库关联）：1996年，美国马萨诸塞州（Massachusetts）发布了医疗患者信息数据库（DB1），去掉患者的姓名和地址信息，仅保留患者的{ZIP, Birthday, Sex, Diagnosis,…}信息。另外有另一个可获得的数据库（DB2），是州选民的登记表，包括选民的{ZIP, Birthday, Sex, Name, Address,…}详细个人信息。攻击者将这两个数据库的同属性段{ ZIP, Birthday, Sex}进行链接操作，可以恢复出大部分选民的医疗健康信息，从而导致选民的医疗隐私数据泄露[3]。

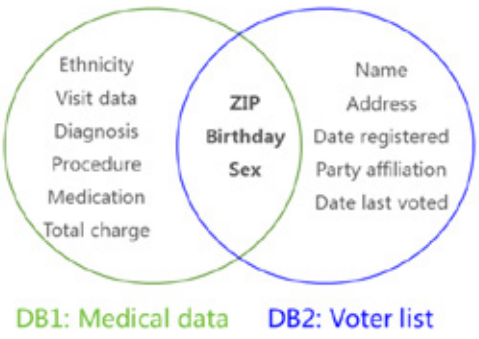


图4 链接攻击（Linking attack）示意图

案例2：AOL导致的用户隐私泄露（语义分析+数据主体关联）： 2006年8月，AOL（美国在线）公司公布了2006年3月1号到5月31号这3个月用户的真实搜

索记录日志，包括1900万搜索，1080多万不一样的搜索词，以及65万余多个经过匿名化处理的用户ID（即将用户注册信息删除，随机ID代替用户真实ID）。虽然用户搜索的日志信息是匿名处理的，然而根据从某个用户ID（随机ID）所做的一系列历史搜索行为和包含信息，仍然有较大可能性分析和关联出用户的真实身份。约时报记者根据搜索数据的地址和姓名等信息找到编号4417749了一位62岁的老太太，家里养了三条狗，患有某种疾病。后面经过老太太本人证实确实是她搜索的关键词。记者曝光该事件后，引起美国公民对AOL公司隐私保护措施的诸多顾虑，并最终导致AOL首席技术官引咎辞职。

以上一些案例说明数据脱敏的仍然会遭受隐私攻击。因此，对经过数据脱敏/匿名化的数据在各种攻击场景的风险评估，或者说对隐私数据脱敏的效果进行检测，尤为重要且关键。

3 数据脱敏效果的评估理论

3.1 研究现状简介

目前，数据脱敏的效果评估技术主要可分为三类：基于人工抽查的定性判定方法、基于模型参数的评估技术和通用的评估技术。其中，基于人工抽查的定性判定方法，指的是按照标准流程和表格进行专家检查和判定，然而，这种方法成本十分昂贵；基于模型参数的评估通常隐私保护模型，包括K-匿名 (K-Anonymity) [3]、L-多样性 (L-Diversity) [4]和ε-差分隐私 (Differential Privacy) [5] 等模型，比如在K-匿名模型中，链接攻击或重识别的最高攻击风险1/K，然而这种指标的刻画能力十分有限，无法充分和全面反映处理后数据的隐私风险分布。通用的风险评估技术与数据脱敏方法与模型无关，在学术上通常称为重识别/重标识攻击（re-identification attack）风险的度量。对于重识别攻击风险问题，最为简单的评估一般使用唯一性 (uniqueness) 指标进行度量，单个属性或者多个属性组合的值在表中是唯一的，那么很可能被攻击唯一识别目标个人信息主体的身份。据权威机构统计，在美国，使用邮编、性别、出生日期信息，有81%的概率可以唯一识别出对应的美国公民。对于大规模数据集的唯一性度量问题， Google学者Chia等人[6] 在2019年的IEEE Symposium on Security and Privacy会议上提出了适应大规模数据集（PB级别）场景的重识别评估算法——KHyperLogLog，他们将重识别问题定义了两种评估指标：Re-identifiability和Joinability，其主要思想是基于Hash、KMV和HLL算法，对结果的近似分析与估计，其在准确率和效率可取得良好的效果。在2019年的Nature communications期刊上，Rocherd等人 [7] 对于发布的数据集重识别评估问题，提出重识别评估的近似模型——Gaussian copulas模型，成功刻画了去标识化/假名化数据集被重识别的成功率，证明即使是不完整的数据集 (比如社会人口调查采样的数据集)，某些个人数据属性的组合仍然存在高的重识别风险。

3.2 重识别攻击场景

在重识别风险评估领域，多个学者进行一系列的指标与评估算法的研究[8-10]，其中较为著名的要数加拿大学者El Emam，他进行了深入的研究，提出了三种常见的攻击场景以及一系列的评估指标 [11]，并将将研究成果进行了转化，创立一家科技公司——Privacy Analytics，主要面向医疗隐私数据，并对去标识化结果进行风险评估与检测，帮助数据处理企业合规美国医疗HIPAA法案，同时获得最大的数据价值，比如将评估合规的脱敏医疗数据出售给保险、药企和科研结构等第三方。

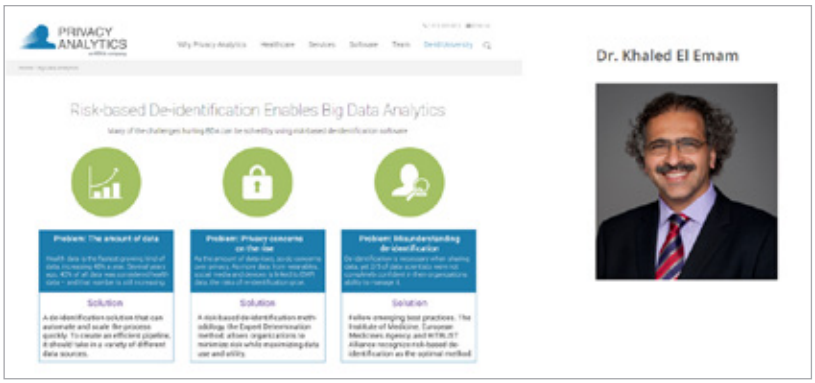


图6 Privacy Analytics 网站主页以及创立者Khaled El Emam (https://privacy-analytics.com)

El Emam根据攻击者的目的和攻击能力定义了三类常见的隐私攻击场景——并形象化地被称为检察官攻击 (Prosecutor attack)、记者攻击 (Journalist attack) 和营销者攻击 (Marketer attack)。其中检察官攻击具有背景知识，了解攻击目标一定在公开数据集中，比如攻击者了解自己的朋友在发布展示的数据集中，他的目的是挑选出他的朋友，并获得敏感属性信息（比如财产消费、医疗健康等信息）；记者攻击场景是达到曝光的目的，他需要尽量寻找公开数据库（比如选举身份登记表），进行匹配关联，进行多次向媒体炫耀，证明重新识别个体，使得涉及企业名誉扫地和难堪；营销者攻击目的是营销，即企业对自己的用户进行多维度的关联与画像，只需保持较高识别概率的匹配关联即可，无需证明是唯一识别出脱敏数据集对应的数据主体。

攻击场景	描述
检察官攻击	攻击者知道某个特定人员在公开的数据集中发生的重标识攻击，他发起的攻击是指向特定目标的，例如同学朋友了解他的同学是受访对象
记者攻击	在此场景中，攻击者一般来说拥有一个大的身份数据库，但他并不知道数据库的人员是否在公开的数据集中，他通过多次炫耀式攻击证明某人可以被重新识别。在这种情况下，攻击者的目标常常是使得公开数据库的组织感到难堪或者名誉扫地
营销者攻击	类似记者攻击场景，但攻击者的目标是使得公开数据库和身份数据库进行关联下实现的重标识攻击。尽量还原出公开数据库的身份，实现精确对身份数据库的人进行其他维度的刻画，但不要求证明重标识结果的正确性，只需保证较高的重标识概率

图7 三种重识别攻击场景

3.3 重识别风险评估指标

检察官攻击、记者攻击和营销者攻击的发生可能性分别检察官攻击风险、记者攻击风险和营销者攻击风险。如何进行定量的隐私风险评估呢？El Emam根据最高、平均重识别风险的概率、高风险记录占有比例等刻画需求在3种场景设计了8类评估指标，具体如图7所示。这些指标可有效地评估结构化的脱敏数据集的隐私风险残余情况。这些指标的数值范围均为[0,1]，1表示最高风险，0表示几乎无风险。

风险类型	评估指标	指标意义	符号含义
检察官攻击风险	$_{prosecutor}\mathcal{R}_{r}=\frac{1}{n}\sum_{j\in J}f_j\times I\left(\frac{1}{f_j}>r\right)$ $_{prosecutor}\mathcal{R}_{m}=\frac{1}{\min_{j\in J}(f_j)}$ $_{prosecutor}\mathcal{R}_{a}=\frac{\lvert J\rvert}{n}$	$_{prosecutor}\mathcal{R}_{r}$ 刻画重识别概率大于 r 的数据集记录占总体的比例； $_{prosecutor}\mathcal{R}_{m}$ 刻画数据集所有记录中最大的重识别概率； $_{prosecutor}\mathcal{R}_{a}$ 刻画平均重识别概率。	①n—数据集记录的数量； ②J—数据集的等价组的集合； ③$\lvert J\rvert$—数据集的等价组数量； ④f_j—数据集等价组为 $j\in J$ 的数量； ⑤r—阈值； ⑥$I()$—当输入为真，输出为1，否则为0； ⑦N—身份数据集记录（可访问或拥有的）的数量； ⑧F_j—身份数据集（可访问或拥有的）等价组为 $j\in J$ 的数量。
记者攻击风险	$_{journalist}\mathcal{R}_{r}=\frac{1}{n}\sum_{j\in J}f_j\times I\left(\frac{1}{F_j}>r\right)$ $_{journalist}\mathcal{R}_{m}=\frac{1}{\min_{j\in J}(F_j)}$ $_{journalist}\mathcal{R}_{a}=\max\left(\frac{\lvert J\rvert}{\sum_{j\in J}F_j},\frac{1}{n}\sum_{j\in J}\frac{f_j}{F_j}\right)$	$_{journalist}\mathcal{R}_{r}$ 刻画重识别概率大于 r 的数据集记录占总体的比例； $_{journalist}\mathcal{R}_{m}$ 刻画数据集所有记录中最大的重识别概率； $_{journalist}\mathcal{R}_{a}$ 刻画平均重识别概率。	
营销者攻击风险	$_{marketer}\mathcal{R}_{a}=\frac{\lvert J\rvert}{N}$ $_{marketer}\mathcal{R}_{1}=\frac{1}{n}\sum_{j\in J}\frac{f_j}{F_j}$	$_{marketer}\mathcal{R}_{1}$、$_{marketer}\mathcal{R}_{a}$ 分别刻画在情况1和2下的平均重识别概率； 情况1：身份数据集和发布数据集的个人信息主体完全相同； 情况2：发布数据集是身份数据集的个人信息主体的一部分。	

图8 三种重识别攻击场景

4 小结

欧盟第29条工作组（Art.29WP）在“Opinion on Anonymisation Techniques 2014”报告[12]中指出匿名化（或称为数据脱敏）“是缓解/降低重识别风险（mitigating the risks）的有效手段”，但并不能完全消除“重识别残余风险”（Residual Risk）。在大数据时代，作为隐私的“防方”——各类脱敏/去标识化/匿名化方法不断更新与发展，作为隐私的“攻方”——去匿名、重识别、“人肉搜索”等攻击与技术也是不断涌现，同时攻击者通过黑客攻击或黑市交易掌握和

收集的用户数据越来越多。因此，数据脱敏/去标识化/匿名化等技术手段处理后的数据集不应被看作是一劳永逸的，数据控制者（Data controller）/数据处理者（Data processor）应该经常性的再评估匿名化后的数据是否存在新的风险。

不仅是欧盟第29条工作组，我国的国标《个人信息去标识化指南》（GB/T 37964-2019）也指出，对数据进行去标识化工作中遵循的一项原则就是，在完成去标识化工作后，应进行评估和定期重新评估。目前，虽然我国多个标准提出要进行数据脱敏效果的评估，然而目前缺乏可落地的流程与技术标准，同时学术界的评估研究几乎为空白，缺乏数据脱敏效果的隐私攻击模型，以及一套统一的评估指标体系，导致国内目前无法评判各类数据厂商脱敏方法的效果与优势。基于此，我们开展国外隐私风险评估理论与技术的研究，并在攻击场景视角下，进一步设计与完善评估模型，以期更加客观地刻画脱敏数据集存在的残余隐私风险。

参考资料

1. Risk Based Security. Data Breach QuickView Report 2019 Q3 trends. 2019-11.

2. 绿盟科技. 2019网络安全观察. http://blog.nsfocus.net/wp-content/uploads/2020/01/2019-Cybersecurity-Insights.pdf

3. Sweeney L, K-anonymity: A model for protecting privacy, International Journal of Uncertainty, Fuzziness and Knowledge-based Systems, 2002, 10(5): 557-570.

4. Machanavajjhala, Ashwin, et al, L-diversity: Privacy beyond k-anonymity. 22nd International Conference on Data Engineering (ICDE'06). IEEE, 2006.

5. Dwork C, “Differential privacy”, Encyclopedia of Cryptography and Security, 2011: 338-340.

6. Chia P H, Desfontaines D, Perera I M, et al. KHyperLogLog: Estimating Reidentifiability and Joinability of Large Data at ScaleIn. IEEE Symposium on Security and Privacy (SP), 2019.

7. Rocher L, Hendrickx J M, De Montjoye Y A. Estimating the success of re-identifications in incomplete datasets using generative models[J]. Nature communications, 2019, 10(1): 1-9.

8. Benitez K, Malin B. Evaluating re-identification risks with respect to the HIPAA privacy rule[J]. Journal of the American Medical Informatics Association, 2010, 17(2): 169-177.

9. Dankar F K, El Emam K, Neisa A, et al. Estimating the re-identification risk of clinical data sets[J]. BMC medical informatics and decision making, 2012, 12(1): 66.

10. Janmey V, Elkin P L. Re-Identification Risk in HIPAA De-Identified Datasets: The MVA Attack[C]//AMIA Annual Symposium Proceedings. American Medical Informatics Association, 2018, 2018: 1329.

11. El Emam K. Guide to the de-identification of personal health information [M]. Auerbach Publications, 2013.

12. Article 29 Data Protection Working Party, Opinion 05/2014 on Anonymisation Techniques,2014.

数据匿名化：隐私合规下，企业打开数据主动权的正确方式?

专题：数据安全
标签：安全合规 隐私保护 大数据
摘要：随着欧盟GDPR、美国CCPA，以及我国《网络安全法》等法规的实施与监管，隐私合规与数据安全治理成为企业当前亟需解决的一大安全任务。具体来说，企业通过技术与管理措施，如何在不影响或少影响原有业务流程的同时去满足合规性？其中，数据匿名化作为一种重要的技术手段，在满足数据统计分析的同时可有效地降低个体隐私泄露风险。且有趣的是，近年来研究发现它具有天然的合规遵循优势。GDPR等法规对赋予用户更多的隐私数据控制权，反过来削减企业的数据控制权与主动权。那么，匿名化技术是否可以帮助企业重新打开数据主动权和控制权这个局面？带着这个疑问，本文将从合规背景、技术算法以及应用与产品三个方面对该技术进行介绍。

1 安全合规背景

欧盟GDPR、美国CCPA赋予了用户非常多的数据权利，例如，GDPR规定用户可对个人数据提出限制处理以及删除的请求；CCPA规定用户有权要求企业不得出售其个人数据。我国《网络安全法》等法规也赋予了一定权利，例如用户发现企业违反规定或错误有权要求企业删除或更正个人信息。

反过来，这些法规对企业提出更高的隐私和安全要求，在一定程度上削弱了企业以往普遍存在的数据掌控能力与权利优势。无疑，这给企业100%的数据掌控权关上了大门，但法规在平衡个体隐私与数据发展的原则指导下，关上一扇门，同时也打开一扇窗——企业可以通过数据匿名化在一些典型数据场景下重新打开数据主动权与控制权。

1) GDPR：对于个人数据、以及假名化等数据，GDPR对相关处理和存储的企业提出十分严厉且全面的法律义务，需要企业履行相关义务。而唯独对于经过处理的匿名数据网开一面——该数据企业可用于统计和研究目的，不受GDPR的约束与限制，即对履行用户各类数据控制权请求等条款具有豁免权。（GDPR前言26段）

2) 《网络安全法》：“匿名”数据（“经过处理无法识别特定个人且不能复原”），企业无需征求被收集者同意，可直接与第三方进行数据共享。（四十二条）

3) 《个人信息安全规范》：个人信息经匿名化处理后所得的信息不属于个人信息（3.14节）；在个人信息主体注销账户场景中，处理注销账户的个人信息有两种方式：① 选择直接删除数据；② 存储匿名化处理后的数据。（8.5节）

由此可看出，匿名化有重要的合规遵循的应用价值，尤其是在数据统计、研究以及数据开放与共享场景中；同时实施该技术措施给企业带来其他方面的好处。即：

1) 合规遵循。匿名数据在向第三方提供、统计分析和注销账户保存匿名数据等场景中是合规的；

2) 数据共享价值。在数据共享场景中，尤其数据敏感且价值密度高的行业，比如医疗，金融等行业，实施数据匿名技术后，可合法合规（光明正大）地进行数据共享与价值挖掘；

3) 增强用户信任。匿名数据，数据是匿名的，即任何人无法识别和关联匿名数据记录的身份。那么用户不担心该数据公开和处理过程中泄露本人隐私；

4) 降低隐私风险。匿名数据在流动和处理过程中，“数据部分可见但身份不可见”，从而有效地降低个体隐私泄露的风险。也就是说，即使匿名数据库遭受黑客攻击外泄，攻击者也无法破解或还原出匿名数据记录所涉及的用户身份信息。

那么什么是匿名化呢？《个人信息安全规范》给出详细的定义：“通过对个人信息的技术处理，使得个人信息主体无法被识别或者关联，且处理后的信息不能被复原的过程”。即匿名化通过数据变换与失真，处理结果可保持一定的可用性，但任何手段无法识别特定个人，且数据不可逆（非“一一映射”（例如加密、置换手段））。

数据脱敏（包括去标识化）作为目前企业广泛实施的数据安全技术，可以看成是“数据匿名化”的相近技术，它对数据进行一系列的数据变换和失真，但无法保证每次处理的结果是真正“匿名”的，即是否达到“无法识别特定个人且不能复原”。若需评估该技术的效果——是否满足法规定义的匿名化门槛，可参考系列文章《数据脱敏后的隐私攻击与风险评估》、《身份证号+手机号如何脱敏才有效？》。如何真正实现和逼近法规的“匿名化”？幸运的是，在学术界中能找到具有广泛和深入研究以K-匿名为代表的数据匿名技术（也称匿名化技术），它可以达到法规要求的匿名化效果。本文下面将对该技术原理、算法，以及现有工业界应用进行介绍，以期进一步促进数据匿名技术在企业场景的研究与应用。

2 数据匿名技术与算法

2.1 概述

早期，个人数据发布的隐私保护场景中，对标识符或准标识符进行简单处理，比如删除、或者使用随机ID替换姓名、用户昵称，对地址信息和出生日期进行泛化处理，这种方式可看成前面提到的 “数据脱敏”。然而随着一些攻击案例和研究发现，这种处理方法的“匿名”处理是不充分的，仍然存在个体隐私泄露的风险。验证这一观点，有多个著名的实际案例：

1) 案例1：1996年美国麻省发布了医疗患者信息数据库（DB1），去掉患者的姓名和地址信息，仅保留患者的{ZIP, Birthday, Sex, Diagnosis,…}信息。另外有另一个可获得的数据库（DB2），是州选民的登记表，包括选民的{ZIP, Birthday, Sex, Name, Address,…}详细个人信息。攻击者将这两个数据库的同属性段{ ZIP, Birthday, Sex}进行关联操作，可以恢复出大部分选民的医疗健康信息，从而一起严重的医疗隐私数据泄露事故。

2) 案例2：AOL公司公布了2006年3个月用户的真实搜索日志，包括1900万搜索记录，为保护隐私对用户ID进行处理，使用随机ID代替真实ID。然而纽约时报记者发现，根据一系列历史搜索行为和包含的相关信息进行推断，可以确定编号4417749的身份——一位62岁的老太太，家里养了三条狗，患有某种疾病。后经过老太太本人证实确实是她搜索的关键词。记者曝光该事件后，引起美国公民对AOL公司隐私保护措施的诸多顾虑，并导致AOL首席技术官引咎辞职。

以上均属于链接攻击（也称重标识攻击、去匿名攻击）范畴，即攻击者通过各种渠道获得公民/用户的身份信息和其他用户的静态属性信息（学术称为“准标识符”属性，比如性别，出生年月、邮编等），包括访问查询公开身份数据集、了解亲朋好友的基本信息、互联网“人肉搜索”陌生人，甚至利用数据泄露、黑灰产数据库等对脱敏数据集进行关联、相似匹配与碰撞，进而还原出上述脱敏数据集的某些记录的身份信息。

为了应对潜在的隐私攻击问题与挑战，学术界开始聚焦和设计隐私保护效果更好的匿名化技术与模型。一般地，用户希望攻击者无法从存在多个个体记录的数据集中识别出自身，以及对应的敏感隐私数据，数据匿名技术便是这种朴素思想的实现之一。Samarati和Sweeney学者在1998年首次提出了匿名化的概念，对个人一些基本信息进行泛化和失真处理，隐藏公开数据记录与特定个人之间的对应联系，从而保护个体的隐私。后面，Sweeney学者在2002年提出了K-匿名模型(K-Anonymity)，该模型保证数据记录的任意等价组至少有K个个体记录，即攻击者无法唯一地确定个体的记录准确身份。如下图所示，它对原始数据进行2-匿名处理，包括对Birth（出生日期）进行泛化、对邮编（ZIP）进行屏蔽处理等操作，最后输出的数据集除敏感属性（Disease）外，其他属性（也称准标识符属性）组成的记录形成等价组，每个等价组至少有两条记录，如索引 (1,2)有2条记录、(2,3) 有2条记录、(4,5) 有2条记录、(5,6) 有2条记录、(7,9) 有3条记录，(10,11) 有2条记录。在攻击场景中，假设攻击者拥有背景知识，了解Jack在该数据集中且掌握了他的基本属性：Race：Black; Birth:1965-09-01；Gender: male；ZIP:02146。攻击者想识别Jack具体属于数据集的那一条记录？经过相似匹配和关联，定位到索引1和索引2，但不能唯一确定那个属于Jack，那么也无法确定Jack患上了那种疾病。也可以说，无法确定索引1和索引2对应的真实身份，从而保护患者的个体隐私。

Index	Race	Birth	Gender	ZIP	Disease
1	Black	1965	male	0214*	short breath
2	Black	1965	male	0214*	chest pain
3	Black	1965	female	0213*	hypertension
4	Black	1965	female	0213*	hypertension
5	Black	1964	female	0213*	obesity
6	Black	1964	female	0213*	chest pain
7	White	1964	male	0213*	chest pain
8	White	1964	male	0213*	obesity
9	White	1964	male	0213*	short breath
10	White	1967	male	0213*	chest pain
11	White	1967	male	0213*	chest pain

图1 经过2-匿名处理的医疗数据

数据发布场景的隐私保护(Privacy Preserving Data Publishing, PPDP)是K-匿名最早的应用，也是研究最为广泛的场景，除此以外将K-匿名为代表的数据匿名技术应用在位置服务和社交网络等领域成为近年来新的一个热点。基于匿名化的PPDP场景可看作为一个通信模型，如图2所示，主要由三方参与：数据控制者/发布者 (Data Controller/Publisher)，可看作发送者；数据接收者(Data Recipients)；隐私攻击者(Attacker)。数据控制者/发布者收集个体 (Individuals) 的个人信息，将这些数据通过匿名化处理 (Data Anonymization) 后得到匿名化数据集，发送给第三方共享或者对外公开。攻击者尝试通过掌握的背景知识和数据库进行攻击，获取具体某个个体的隐私信息。典型一种攻击方式是链接攻击，即去除准标识符信息 (Identifier，ID，如姓名，身份ID)，攻击者通过其他渠道掌握的数据库的同属性段(称为准标识符，Quasi-Identifier, QID)与公开数据库进行链接和匹配操作，恢复出具体个体敏感信息(Sensitive attribute, SA，如健康、薪资、位置等)。

图2 数据匿名化的一般应用场景

2.2 模型与算法

数据匿名技术的研究主要集中在模型、算法、匿名处理操作和评估指标四个研究方面。

2.2.1 匿名模型

前面提到的K-anonymity由于敏感属性进行约束，当等价组的敏感属性取值相同时，仍然存在隐私风险，Machanvajjhala等人提出了L-diversity模型，在每一个等价组中，至少存在L个不同的敏感属性，相比K-anonymity增强了安全性。Li等人在L-diversity基础上，考虑敏感属性分布，设计了T-closeness模型，通过保证任意等价组的敏感属性的分布与敏感属性的全局分布之间的距离小于T，进一步增强了安全性，然而约束条件越来越多，降低数据的可用性。除以上模型外，还出发展和衍生出了(α,k)-anonymity和个性化隐私保护 (Personalized privacy

preservation)的模型等。

2.2.2 匿名化算法

匿名化算法以最小的数据缺损代价实现满足模型的约束。然而研究表明，实现最优的匿名化是一个NP难题。幸运的是，目前已经发展许多有效的近似算法，典型的算法包括Datafly算法，Mondrian算法。前者是单维度泛化算法，其核心思想是对给定数据表中QID的属性中取值最多的那个属性按预先给定的泛化树进行泛化，直到匿名化数据表满足K-anonymity；后者是多维度泛化算法，其核心思想是将所有QID属性看成是一样的，即只有一个等价组，然后自上而下，启发式选择QID的某个属性逐次划分，直到满足条件无法划分。除以上算法外，由于聚类算法思想与匿名化等价类划分思想十分相近，因此一些学者提出基于聚类的匿名化算法。

2.2.3 匿名处理操作

主要包括数泛化、抑制、置换等操作。其中泛化最为广泛应用，泛化是指用模糊/抽象/概括的值代替精确值，使得多个数据是相同的。例如年龄26，29被泛化为“25-30”，地址朝阳区、海淀区被泛化为北京市，那么攻击者无法精确地获得数据主体精确信息；抑制操作一般将数据使用“*”代替，隐藏和遮蔽数据值，使得攻击者无法获得该部分的信息；置换是对数据表中的属性值进行位置打乱操作，使得数据主体与该属性信息不对应，一般用于SA属性的处理中。

2.2.4 评估指标

主要分为两个方面的评价，数据可用性（Data Utility）与隐私保护性(Privacy Protection)。在文献研究中，前者指标较为丰富，可对，包括匿名化的NCP (Normalized Certainty Penalty)，CM (Classification Metric) 和DM (Discemibility Metric)。后者研究文献，一般默认使用模型的参数进行评判，例如K-anonymity、L-diversity，参数K和L越大，分别对应的重识别和隐私泄露风险越小。近年来，一些学者基于通信模型和Shannon信息论对，对隐私泄露问题进行数学建模与分析，提供了理论的度量方法。

3 数据匿名技术的应用

数据匿名技术随着发展逐步趋向成熟，一些高校和研究机构基于软件功能实现开源数据匿名化项目与工具，一些面向隐私合规的欧美科技公司对该技术进行

产品化和应用。

3.1 开源项目

基于数据匿名技术的工具化实现主要集中在欧美高校和研究结构，有4个著名的开源项目：ARX、UTD Anonymization Toolbox、Cornell Anonymization Toolkit或Amnesia。从成熟度看，ARX最为成熟，提供丰富的界面和API接口，以及在微软匿名化，提供完整的数据可用性、重标识风险评估等功能组件。

表1 数据匿名的相关开源项目

	ARX	UTD Anonymization Toolbox	Cornell Anonymization Toolkit	Amnesia
开发者机构	慕尼黑工业大学·德国	得克萨斯大学达拉斯分校·美国	康乃尔大学·美国	信息系统管理研究所（IMSI）·希腊
开发语言	Java	Java	C++	Java
项目主页 / github	https://arx.deidentifier.org https://github.com/arx-deidentifier/arx	http://cs.utdallas.edu/dspl/cgi-bin/toolbox	https://github.com/wanghaisheng/Cornell-Anonymization-Toolkit	https://amnesia.openaire.eu https://github.com/dTsitsigkos/Amnesia
项目成熟度	实验研究，半产品	实验研究	实验研究	实验研究，接近半产品
支持的匿名模型	K- 匿名、L- 多样和 T- 近似	K- 匿名、L- 多样和 T- 近似	L- 多样	K- 匿名、K^m- 匿名
支持的匿名算法	Flash	Datafly、Mondrian、Incognito	Incognito	Flash、基于聚类算法
特点	提供丰富的数据可用性、风险评估等功能	提供多种匿名算法实现	可简单进行数据可用性、风险评估的计算	支持在线 https://amnesia.openaire. eu/amnesia

3.2 企业产品应用

GDPR、CCPA的隐私合规驱动，一些欧美企业，包括Google，以及主打隐私合规产品的创业公司，率先将数据匿名技术进行了孵化与产品应用。

1) Google 的云 DLP 产品

Google 浏览器的隐私声明中，承诺对用户数据使用K-匿名、L-多样数据匿名化以及差分隐私等技术进行处理。随着用户隐私与敏感数据上云，隐私和数据泄露问题引起云使用者的担忧，谷歌将匿名化技术嵌入DLP产品中，可以解决隐私风险问题。DLP产品实现四种匿名化模型与算法，包括K-匿名、L-多样、K-图和σ-

存在性，用户可以根据隐私保护和数据统计分析的需求选择合适的模型算法。在匿名处理数据前，云DLP谷歌提供了原始数据的风险洞察功能：如图3所示，使用者可以看到在K-匿名的不同K值下，不满足的记录数比例（蓝色）。

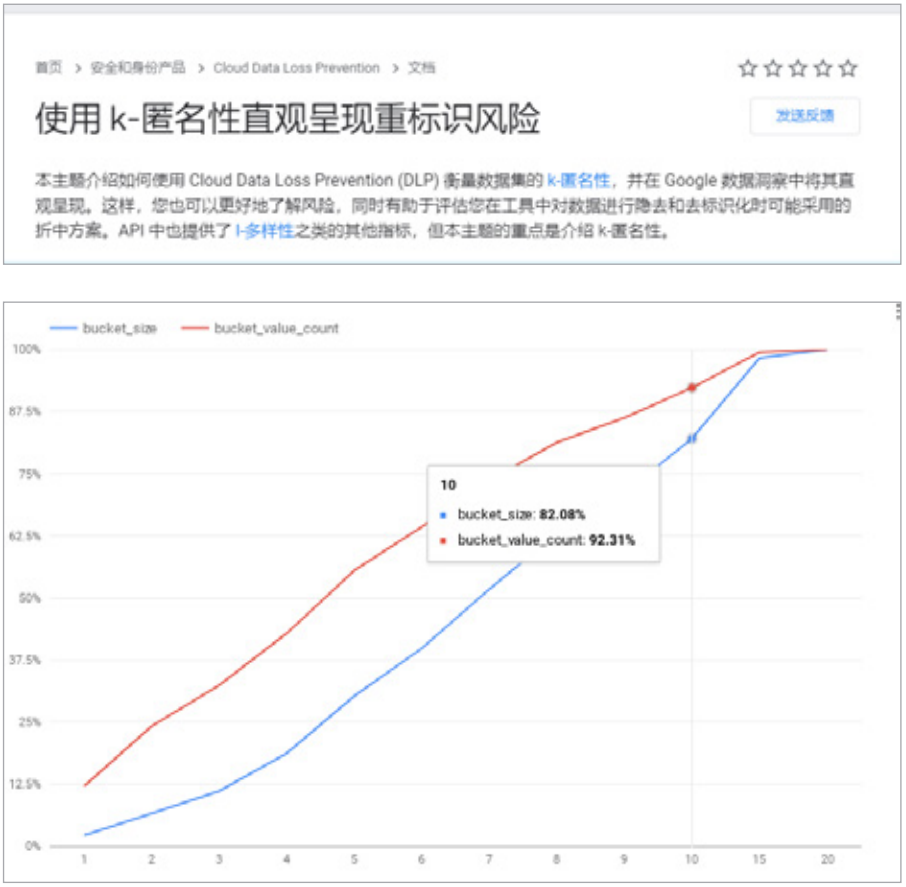


图3 Google云DLP产品的风险洞察功能

2) Immuta 的数据治理平台产品

Immuta是一家美国初创企业，目前处于C轮融资（融资总额6820万美元）。Immuta在云云生数据治理平台 (cloud-native data governance platform) 应用到了K-匿名技术，K-匿名可应用在静态数据和动态数据中，后者可能采用类似m-invariance的匿名算法，即保持动态增量的数据仍然满足等价组数量至少为K个，该技术的应用可增强云存储与计算的隐私安全。

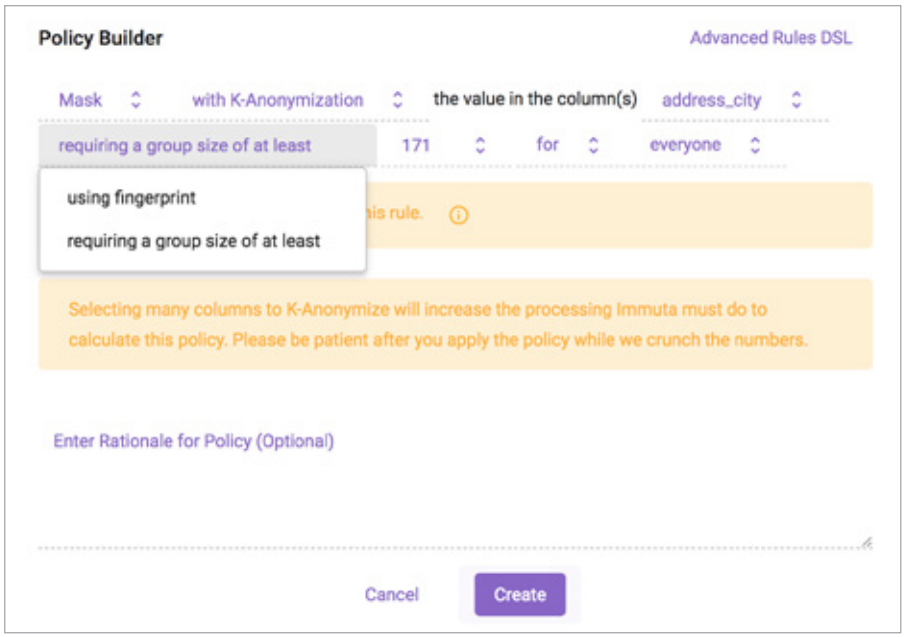


图4 Immuta的数据治理平台

3) Privitar 的数据脱敏产品

Privitar是一家成立于2014年总部位于英国伦敦的创业公司，目前处于C轮融资（融资总额15050万美元），主打推出一系列的隐私产品，包括大规模数据隐私治理的自动化、隐私政策管理、数字水印平台以及数据脱敏产品(公司称为De-Identification 产品，实际功能与国内的Data masking功能基本相同)。在数据脱敏系统中，除了使用传统的基于泛化、替换、屏蔽和加密等脱敏策略外；其数据脱敏嵌入了K-匿名算法，它相比传统脱敏的策略在隐私保护上更强优势，攻击者即使获得经过K-匿名的脱敏数据，也无法通过其他渠道获得的身份信息或数据库进行关联推断，还原脱敏记录的真实身份，进而有效保证隐私前提下实现脱敏数据的提取与利用。

4) Anonos 的 BigPrivacy 产品

Anonos是一家美国初创企业，目前融资总额1200万美元。其公司主要推出了BigPrivacy产品，同样可以看成一个脱敏平台，其假名化和身份与业务数据分离，这些功能可满足GDPR具体条款的一些合规性。在该平台中，Anonos也应用了K-匿名技术与算法，在保证数据在业务场景的可用性时，可保证K-匿名处理后的数据不被重标识与身份识别。

4 小结

全球的数据安全隐私法规的立法，一方面赋予了公民与互联网用户的数据控制权利；另一方面对数据处理的企业提出了更高的隐私与安全要求。企业一方面可以通过访问控制和网络安全防护等措施降低数据收集、存储和处理等阶段的隐私泄露风险，另一方面在日益增多的数据共享与计算场景实施数据匿名化是不错的选择——不仅满足业务利用与隐私保护，同时遵循了合规性。

参考资料

1. Samarati P, Sweeney L. Generalizing data to provide anonymity when disclosing information (abstract). symposium on principles of database systems, 1998.

2. Sweeney L. K-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-based Systems, 2002,10(5):557-570

3. L-diversity:Pri-vacy beyondk-anonymity. Machanavajjhala A,Gehrke J,Kifer D,et al. Proceedings of the 22th International Conference on Data Engineering . 2006

4. Li N H, Li T C, Venkatasubramanian S. T-Closeness-privacy beyond K-anonymity and L-diversity. IEEE 23rd International Conference on Data Engineering, Istanbul, Turkey, April 15-20, 2007: 106-115.

5. Dwork C. Differential privacy. Encyclopedia of Cryptography and Security, 2011: 338-340.

6. Rocher L, Hendrickx J M, De Montjoye Y A. Estimating the success of re-identifications in incomplete datasets using generative models. Nature communications, 2019, 10(1): 1-9.

7. El Emam K. Guide to the de-identification of personal health information. Auerbach Publications, 2013.

8. Meyerson, Adam, and Ryan Williams. "On the complexity of optimal k-anonymity", Proceedings of the twenty-third symposium on Principles of database systems. ACM, 2004.

9. LeFevre, K., D. J. DeWitt, and R. Ramakrishnan. "Mondrian Multidimensional K-Anonymity" 22nd International Conference on Data Engineering (ICDE'06). IEEE, 2006.

10. Aggarwal, Charu C, "On k-anonymity and the curse of dimensionality", Proceedings of the 31st international conference on Very large data bases. VLDB Endowment, 2005.

11. Zakerzadeh H, Aggarwal C C, Barker K, "Towards breaking the curse of dimensionality for high-dimensional privacy", Proceedings of the 2014 SIAM International Conference on Data Mining, 2014: 731-739.

12. Ayala-Rivera, Vanessa, et al, "A systematic comparison and evaluation of k-anonymization algorithms for practitioners", Transactions on data privacy 7.3 (2014): 337-370.

隐私保护与价值挖掘之利器——数据脱敏、匿名化、差分隐私与同态加密

古人云，“鱼，我所欲也，熊掌亦我所欲也；二者不可得兼”。大数据时代，数据挖掘诚可贵，例如各类APP通过收集我们的行为信息进行购买商品与美食预测和推荐，提高用户体验和提升效率；然而，隐私保护价更高，例如敏感的个人信息（姓名、家庭住址和手机号码等）被某些机构收集，为了某种利益被非法贩卖或被黑客攻击泄露，定向电信诈骗由此而生，山东徐玉玉案件给社会敲响了警钟。在大数据的应用场景下，在满足数据安全和隐私保护的同时，实现数据的流动和价值的最大化/最优化成为“数据控制者”或“数据处理者”普遍诉求。幸运的是，经过信息技术的发展和革新，“鱼和熊掌兼得”成为可能：数据处理者/控制者不但能收获到那条“鱼”（价值挖掘），也能得到预想的那只“熊掌”（隐私保护）。

在该系列的第一篇中介绍了国内外的数据安全与隐私保护相关法规，如欧盟《GDPR》、美国《CCPA》和中国《网安法》。这些法规保护的个人数据(或个人信息)范畴均十分广泛，且具有严格的约束和规范。在法规指导下，如何更好地满足合规，降低法律风险和隐私泄露风险；同时也能满足业务场景需求。目前存在多种关键技术，场景不同，需求不同，对应的技术也自然不同。本文作为《大数据时代下的数据安全》系列的第二篇：场景技术篇，将介绍四种关键技术：数据脱敏、匿名化和差分隐私和同态加密，并对每一种介绍技术的从场景、需求和技术原理等几个维度进行展开。

1 数据脱敏

数据脱敏，也称为数据漂白（英文称为Data Masking或Data Desensitization）。由于其处理高效且应用灵活等优点，是目前工业界处理敏感类数据（个人信息，企业运营、交易等敏感数据）普遍采用的一种技术，在金融、运营商、企业等有广泛应用。广义地讲，人脸图像打码（马赛克）实际也是一种图片脱敏技术：通过部分的屏蔽和模糊化处理以保护“自然人”的隐私。但本文讨论的是传统的（狭义的）脱敏技术——即数据库（结构化数据）的脱敏。

场景

数据库是企业存储、组织以及管理数据的主要方式。几乎所有的业务场景都与数据库或多或少有所关联。在高频访问、查询、处理和计算的复杂环境中，如何保障敏感信息和隐私数据的安全性是关键性问题。对于个人信息使用和处理场景，主要有以测试、培训、数据对外发布、数据分析等为目的

场景。举一个测试场景例子。假如小明是测试人员，在进行产品测试过程中，需要使用一些用户个人信息示例数据。如果可以直接访问和下载用户个人信息的原始数据，那么有隐私泄露的风险（他可能将用户个人信息卖给另一家公司）。为了避免风险，可对所有数据项逐一进行加密。但这引起了一个问题——数据的密文数据杂乱无章，已经失去了测试和验证价值。那么是否可以在数据可用性（Data Utility）和隐私保密性（Privacy Protection）进行折中呢，答案是肯定的。4. 如示意图1中，小明需要访问用户信息数据库，服务器根据小明的权限对数据颗粒度进行管控和脱敏处理，比如仅保留姓、年龄进行模糊处理（四舍五入）、电话号码屏蔽中间四位。那么小明无法得到准确无误的用户信息，或者猜测次数过多（猜测概率过低）带来的攻击成本不足以支撑小明的攻击动机（铤而走险）。

图1 测试场景：使用脱敏数据

需求

个人信息或其他敏感信息的处理，必须满足两个要求：

1) 数据保密性（Data confidentiality）：对于个人信息，称为隐私保密性（Privacy Protection）。需要保证潜在的攻击者无法逆推出准确的敏感信息，对于一些关键信息无法获取。

2) 数据可用性（Data Utility）：保证被处理后的数据，仍然保持某些统计特性或可分辨性，在某些业务场景中是可用的。

这两个指标是一对矛盾。如何调节与平衡：哪些数据字段需要加强保密？哪些字段可以暴露更多信息？屏蔽多少信息可达安全/应用？这些需要分析和研究具

体应用场景，再一步细化两个指标需求（场景需求的定制化）。比如某一个APP的业务场景，需要统计和分析APP用户的年龄分布，为了保护用户隐私，需进行处理和失真，但需尽可能保留年龄字段的统计分布。如何达到呢？下面即将介绍。

技术原理

数据脱敏是解决上述需求的关键技术。所谓脱敏，是对敏感数据通过替换、失真等变换降低数据的敏感度，同时保留一定的可用性、统计性特征。为了达到这个目标，有一系列的方法/策略可以使用。以下表格列举一些典型的脱敏方法/策略。具体使用哪种脱敏方法，需要根据业务场景，如数据的使用目的、以及脱敏级别等需求去选择和调整。如上述统计APP用户年龄分布的例子，可使用重排的方法保证数据的统计分布。

方法/策略	描述	示例
取整	数值或日期数据的取整	13:25:15 → 13:00:00
量化	通过量化间距调整数据失真程度	27 → 30
屏蔽	屏蔽部分数据，如电话、身份证号码	152****1234
截断	数据尾部截断	010-88886666 → 010
唯一替换	使用替换表对敏感数据进行替换	231→1 20→2 231→1
哈希	将输入映射为固定长度的字符串	8 → a17d 28 → 1c4a
重排	将数据库的某一列值进行重排	22,31,27→ 31,27,22
FPE加密	明文和密文格式不变，属于统一集合	15266661234 → 15173459527

图2 常见脱敏方法/策略

其中，保留格式加密（Format-Preserving Encryption，FPE）是一种特殊的加密方式，其输出的密文格式仍然与明文相同。比如中国联通手机号15266661234，通过FPE加密可以实现仍然输出的是联通手机号15173459527。 FPE加密应用时，需考虑格式及分段约束，这与一般的对称分组加密不同。为了规范FPE技术实施，美国NIST发布了FF1标准算法，可用于保险号、银行卡号、社保卡号等数字标识符的加密与脱敏。

应用

按照使用场景，可将脱敏分为静态脱敏(Static Data Masking, SDM)、动态脱敏(Dynamic Data Masking, DDM)，本文介绍是前者。静态脱敏一般用于非生产环境中（测试、统计分析等），当敏感数据从生产环境转移到非生产环境时，这些原始数据需要进行统一的脱敏处理，然后可以直接使用这些脱敏数据；动态脱敏一般用于生产环境中，在访问敏感数据当时进行脱敏，根据访问需求和用户权限进行“更小颗粒度”的管控和脱敏。一般来说，动态脱敏实现更为复杂。脱敏在多个安全公司已经实现了应用，IBM，Informatica公司是比较著名的代表。

2 匿名化

匿名化技术（Anonymization）可以实现个人信息记录的匿名，理想情况下无法识别到具体的“自然人”。主要有两个应用方向：个人信息的数据库发布或挖掘（Privacy Preserving Data Publishing，PPDP，或Privacy Preserving Data Mining，PPDM）。

场景

一个经典的场景，是医疗信息公开场景。医疗信息涉及患者个人信息以及疾病隐私，十分敏感；但对于保险行业的定价、以及数学科学家对疾病因素等各项研究，这些数据具有巨大的价值所在。为了保护患者的身份和隐私，让人很容易想到的是删除身份有关信息,即去标识化(De-identification)。关于此，一个经典案例，美国马萨诸塞州发布了医

疗患者信息数据库（DB1），去掉患者的姓名和地址信息，仅保留患者的{ZIP, Birthday, Sex, Diagnosis,…}信息。另外有另一个可获得的数据库（DB2），是州选民的登记表，包括选民的{ZIP, Birthday, Sex, Name, Address,…}详细个人信息。攻击者将这两个数据库的同属性段{ ZIP, Birthday, Sex}进行链接和匹配操作，可以恢复出大部分选民的医疗健康信息，从而导致选民的医疗隐私数据被泄露[1]。

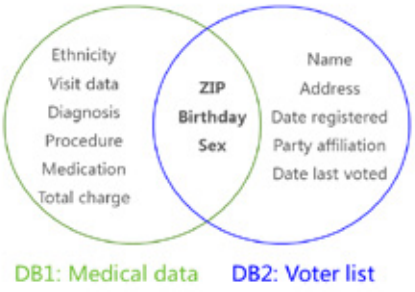


图3 链接攻击：两个数据库的链接

需求

1）无法重识别（De-identification）：通俗地讲，如何使得发布数据库的任意一条记录的隐私属性（疾病记录、薪资等）不能对应到某一个“自然人”，无法实现“重识别”，即切断“自然人”身份属性与隐私属性的关联。

2）数据可用性（Data Utility）：尽可能保留数据的使用价值，最小化数据失真程度，满足一些基本或复杂的数据分析与挖掘。

技术原理

为了满足以上需求，一般使用匿名化技术(Anonymization)。在学术研究上，最早由美国学者Sweeney提出，设计了K匿名化模型(K-Anonymity)[1]。即通过对个人信息数据库的匿名化处理，可以使得除隐私属性外，其他属性组合相同的值至少有K个记录。为了展示匿名化过程，下图给出了关于薪资的个人信息匿名化的例子。

图4 2-匿名化示例：保护薪资隐私信息

对于大尺寸的数据表，如何实现K-匿名化的目标呢？这是算法实现和复杂度优化的问题，目前有基于泛化树和基于聚类的匿名化实现方法。除K-匿名化外，还发展和衍生出了(α, k)-匿名 ((α, k)–Anonymity)[2]、L-多样性(L-Diversity)[3] 和T-接近性 (T-closeness)[4]模型。在具体应用时，需要根据业务场景（隐私保护程度和数据使用目的）进行选择。

概念辨析

需辨别的是，匿名化(Anonymization)、假名化(Pseudonymization)、去标识化(De-identification)三个概念有些联系，但不尽相同，却常常被混为一谈。

假名化(Pseudonymization)：将身份属性的值重新命名，如将数据库的名字属性值通过一个姓名表进行映射，通常这个过程是可逆。该方法可以基本完好保存个人数据的属性，但重识别风险非常高。一般需要通过法规、协议等进行约束不合规行为保证隐私的安全性。

去标识化(De-identification)：将一些直接标识符删除，如上表所示，去掉身份证号、姓名和手机号等标识符，从而降低重识别可能性。严格来说，

根据攻击者的能力，仍然有潜在的重识别风险，见图3。

匿名化(Anonymization)：通过匿名化处理，攻击者无法实现“重识别”数据库的某一条个人信息记录对应的人，即切断“自然人”身份属性与隐私属性的关联。

一般来说，这三种方法对数据可用性依次降低，但隐私保密性越来越高。

3 差分隐私

差分隐私(Differential Privacy, DP)具有严格的数学模型，无需先验知识的假设，安全性级别可量化可证明。是近年来学术界隐私保护研究热点之一，同时，一些企业应用将差分隐私技术应用到数据采集场景中。

场景

一个典型的场景：统计数据库开放，比如某家医院提供医疗信息统计数据接口，某一天张三去医院看病，攻击者在张三去之前(第一次)查询统计数据接口，显示糖尿病患者是人数是99人，去之后攻击者再次查询，显示糖尿病患者是100人。那么攻击者推断，张三一定患病。该例子应用到了背景（先验）知识和差分攻击思想。

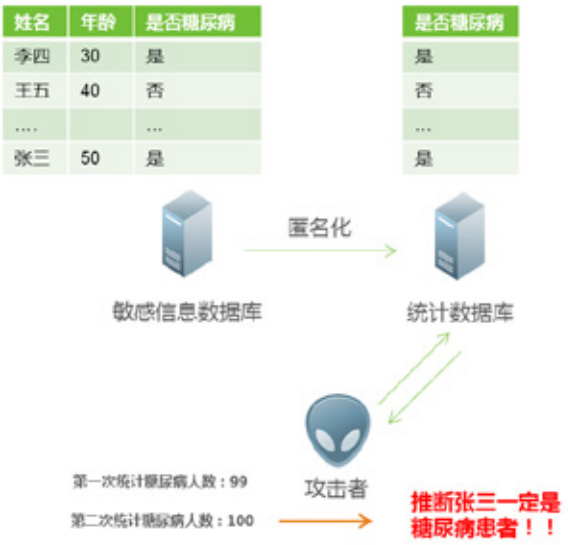


图5 攻击场景：应用背景知识和差分攻击获取隐私信息

需求

上述场景要求设计一种算法：即使攻击者拥有一定背景知识（先验知识），攻击者查询公开数据库，只能获得全局统计信息（可能存在一定误差），无法精确到某一个具体的记录（“自然人”的记录）

技术原理

为了这个需求，差分隐私技术(Differential Privacy, DP)应运而生。这项技术最早由微软研究者Dwork 在2011年提出[5]。DP可以确保数据库插入或删除一条记录不会对查询或统计结果造成显著影响，数学化描述如下：

$$\frac{\Pr(f(D)=C)}{\Pr(f(D')=C)}<e^{\varepsilon}$$

D和D'分别指相邻的数据集（差别只有一条记录），f(.)是某种DP算法，它对于任意的输出C，两个数据集输出的概率几乎接近（小于e^ε）那么称为满足ε隐私。如何实现这个目标呢？一般来说，通过在查询结果加入噪声（如Laplace噪声），使得查询结果在一定范围内失真，并且保持两个相邻数据库概率分布几乎相同。那么DP方法可以抵抗差分攻击引起的隐私泄露。比如上述场景，第一次查询结果是99个，第二次查询概率为 p 结果为99个，1-p 的概率结果是100个，那么攻击者无法准确地确认张三是否患病。

应用

以上介绍的是中心化的差分隐私 (Centralized Differential Privacy ,CDP)。随着研究的进展，出现了本地差分隐私(Local Differential Privacy, LDP)。LDP在用户侧进行，服务器无法获得真实的隐私信息，其核心思想是随机化算法，即每一个采集的数据都加入了噪声。若采集的数据足够多，那么得到相对准确的统计分布。LDP的原理注定它十分适用于用户隐私数据的采集。一些IT公司开始应用该项技术，比如iPhone使用LDP技术用户隐私，在可获得统计行为的同时，避免用户隐私的泄露。如图3是iPhone手机提供的说明例子（感兴趣读者可去手机查找和研究）[6]，每一个用户的表情加入了噪声，是不准确的，但经过大量用户的频率统计，是相对准确的；Google也进行类似的应用，通过Chrome浏览器使用LDP技术采集用户行为统计数据。

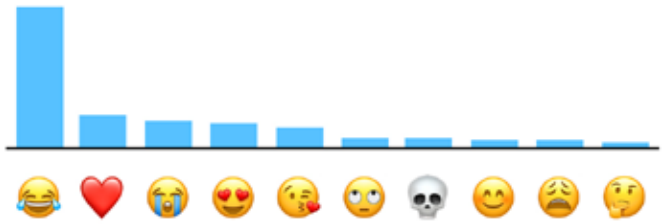


图6 iPhone使用本地差分隐私技术：采集用户表情信息

4 同态加密

同态加密不同于传统的加密，它是应对新的安全场景出现的一项新型密码技术。它的出现，颠覆了人们对密码算法认知。使得密文处理和操作，包括检索、统计、甚至AI任务都成为可能。

场景

假设创业公司C拥有一批数据量大且夹杂个人信息的数据，需要多方进行共享和处理。为了降低成本，他选择使用廉价的不可信第三方平台：公有云。但为了保障传输和存储过程的数据安全，公司员工C1在数据上传前，对

数据进行了加密，再将得到的密文数据上传到共有云。公司员工C2，需在共有云上执行一个数据分析和统计的任务。

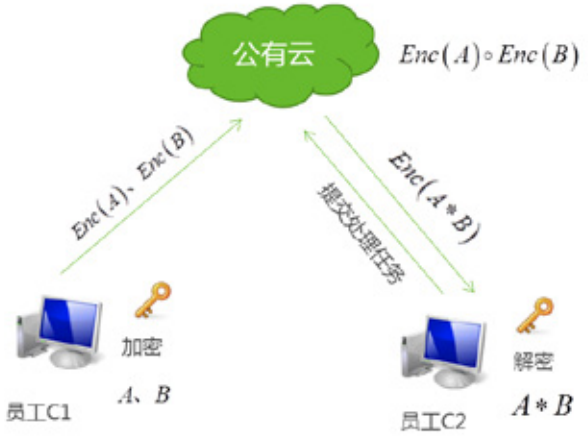


图7 云平台的安全计算场景

需求

以上的场景可提炼出两个需求：

1. 安全需求：除了公司C员工可解密数据外，其他人包括第三方平台无法解密和查看数据，即需要保障个人隐私数据的安全性；

2. 处理需求：存储在第三方平台的密文数据，仍然可以进行基本运算（加减乘除）、统计、分析和检索等操作。处理后的密文数据，返回给公司C的员工，得到结果和预期是一致的。

技术原理

同态加密满足上述需求的一项关键的技术之一。假设A,B是两个明文，Enc(.)是加密函数，那么其存在以下性质，

$$Enc(A)°Enc(B)=Enc(A*B)$$

该性质在数学上称为同态性。通俗地讲，在密文域进行 操作相当于在明文域进行 操作。这种性质使得密文域的数据处理、分析或检索等成为可能。如下示意图所示，假设员工C1上传两个密文数据Enc(A)、Enc(B)（对应两个明文数据为为）到不可信的云平台中，员工C2提交两个明文数据 的任务，那么云计算平台对应执行密文的操作是：Enc(A)°Enc(B)。从始至终，云平台一直没有接触到相关的明文信息，从而防止了第三方窃取导致的隐私数据泄露。

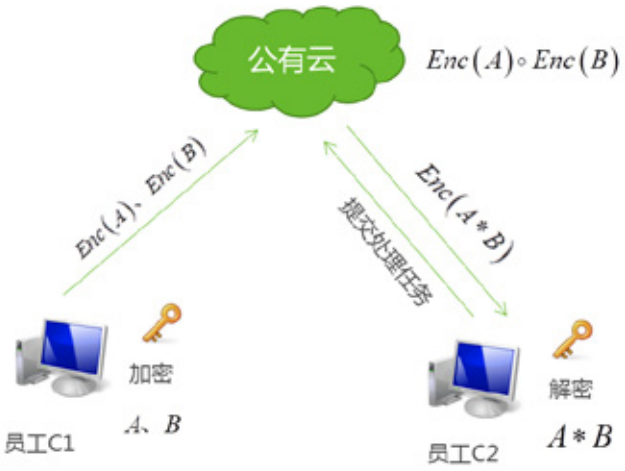


图8 同态加密在云平台的应用

应用

同态加密过程需要消耗大量的计算资源。但目前开始有一些开始朝向应用发展：同态加密逐步开展了标准化进程；另外创业公司Duality在定制服务器通过同态加密，实现隐私保护与AI任务等应用。

5 小结

大数据时代，隐私保护诚可贵，数据挖掘价更高。根据实际应用场景，处理和平衡数据可用性(Data Utility) 和隐私保密性(Privacy Protection)，是大数据时代下的数据安全的关键性问题之一。在保留一定的数据可用性、统计性等基础上，通过失真等变换实现降低数据敏感度——数据脱敏；通过“去识别化”实现隐私保护——匿名化；6.　　通过加噪来抵抗差分攻击——差分隐私；甚至将个人敏感信息直接加密，然后在密文数据上进行统计与机器学习——同态加密。然而，其中一些技术在具体的场景落地时，仍然面临着诸多挑战，如数据可用性和隐私保护如何实现自适应调节；高维、大数据集的效率问题如何优化等是值得深入研究的问题。

参考资料

1. Sweeney L. K-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-based Systems, 2002,10(5):557-570.

2. Wong R C, Li J, Fu A W, et al. (α, k)-anonymity: An enhanced kanonymity model for privacy-preserving data publishing [C]. The 12thACM SIGKDD International Conference on Knowledge Discovery andData Mining, Philadelphia, PA, USA, August 20-23, 2006.

3. l-diversity:Pri-vacy beyondk-anonymity. Machanavajjhala A,Gehrke J,Kifer D,et al. Proceedings of the 22th International Conference on Data Engineering . 2006

4. Li N H, Li T C, Venkatasubramanian S. t-Closeness-privacy beyond kanonymity and l -diversity [C]. IEEE 23rd International Conference on Data Engineering, Istanbul, Turkey, April 15-20, 2007: 106-115.

5. Dwork C. Differential privacy[J]. Encyclopedia of Cryptography and Security, 2011: 338-340.

6. https://www.apple.com/privacy/docs/Differential_Privacy_Overview

NSFOCUS

安全观察

透过隐私合规，看数据安全技术发展趋势

专题：数据安全
标签：GDPR、个人信息保护法（草案）、Gartner成熟度曲线、数据安全技术
摘要：近年来，全球掀起个人信息与隐私的立法热潮。欧盟2018实施GDPR，美国2020年实施CCPA，两部法规均对企业处理用户的数据提出更严、更具体的约束和要求；最近十月份，我国对外公布《个人信息保护法（草案）》，它全面和具体地规定了企业保护个人信息安全的各项义务,同时指出违反法规最高可面临5000万或一年度营业额5%的巨额罚款。

据Gartner预测，到2023年年底，全球超过80%的企业将面临至少一项隐私数据保护的法规（跨国企业面临多个国家或地区的多项隐私法规）。在法规监管不断强化的背景下，企业不得不重新审视数据安全与合规性的重要性与急迫性。与此同时，数据安全技术近年来发展十分迅速，创新技术不断涌现。本文将从国内外隐私合规视角切入，对数据安全技术进行梳理和总结，并对国内外数据安全技术发展趋势进行洞察和分析。

1 监管不断强化的国内外隐私法规

2018年 5月25日，欧盟正式实施《通用数据保护条例》（General Data Protection Regulation，GDPR）[1]，取代了1995年起施行的《数据保护指令》。GDPR不仅保护欧盟境内的个人数据，以及境外的欧盟公民的个人数据（域外管辖权）。GDPR赋予数据主体（用户）更多的数据控制权：不仅包括原有法规的知情权、访问权、修改权等，同时增加“被遗忘权”和“可携带权”两项“特权”。被遗忘权，在一些注销账户、或者超过时间期限等场景中，用户可以行使

该项权利——数据控制者（企业）收到权利请求后，允许删除与自己相关的个人数据，同时需要通知合作的第三方也删除相关的个人数据；可携带权，用户可以便携地将其个人数据从一个数据控制者处转移至另一个数据控制者处，数据控制者需要配合完成该过程。同时，GDPR规定企业保护数据需采取假名化、加密以及其他技术措施，数据泄露采取快速响应机制等等。此外，违法的代价是高昂的——最高罚款额度在2000万欧元或公司全球营业额的4%。从2018年执法到现在，多数成员国已经陆续开出多张的罚单。非常具代表性的一家大型国际互联网公司——Google在隐私保护方面已经做了不少工作，然而Google却陆续被欧盟的两个国家罚款：2019年1月份被法国处罚5000万欧元，原因是执法方认为Google产品的隐私条款未充分体现GDPR公开透明和清晰原则；2020年3月被瑞典处罚700万欧元，原因是Google未能充分履行GDPR赋予用户的数据“遗忘权”。

受GDPR立法的影响，全球其他国家也陆续推出了相关的隐私法规。具有代表性的是美国2018年6月通过的《加州消费者隐私法案》（California Consumer Privacy Act，CCPA），由于影响涉及大部分知名IT科技公司，如惠普、Oracle、Apple、Google和Facebook等，该方案从立法到颁布备受各界人士的关注。该法规同样赋予了消费者多种数据权利，同时对企业提出更严的标准与要求。另外，巴西于2019年7月通过《通用数据保护法》（LGPD）的最终版本；印度在2018年12月公布修改后的《2019年个人数据保护法（草案）》（Personal Data Protection Bill, 2019）；泰国于2020年5月正式实施《个人数据保护法》（Personal Data Protection Act）等。

2020年10月21日，我国《个人信息保护法（草案）》在人大网正式对外公布[2]。作为一部全面保护个人信息安全的综合性法律，具有重要的意义。该法律保护我国境内公民的各项个人信息权益，同时赋予个人信息主体各项数据权利，包括知情权、决定权、查询权、更正权、删除权等；同时明确了个人信息处理者（企业）的合规管理和保障个人信息安全等义务，并指出保障个人信息安全采取分级分类、加密、去标识化等措施。此外，对违法的行为提出更高的处罚力度，违反法规最高面临5000万元人民币或一年度营业额5%的巨额罚款，同时可以责令暂停相关业务、停业整顿、吊销营业许可或营业执照等严厉的行政处罚。这些处罚给企业的个人信息违规违法行为形成强大的威慑力。值得关注的是，在该草案公布临近几天，金融领域执法重拳出击：央行对3家银行的6家分支机构由于侵害消费者个人信息等违规行为开出百万、千万级大额罚单，并对相关责任人予以警告并处以罚款[3]。可见，企业应足够重视个人信息安全与数据隐私合规性问题，并落实相关举措。

从对企业的影响来看，对欧盟GDPR和国内的《个人信息保护法（草案）》以下的一些合规性热点进行解读：

1) 个人数据/个人信息的识别与分类

GDPR保护的数据对象是 “个人数据”。其定义是“关于一个已识别或者可能识别的自然人（即数据主体）的任何信息”，“个人数据”范畴边界十分宽泛，涵盖信息十分丰富，不仅包括传统意义的姓名、年龄、性别这些基本的个人信息，还包括一些特殊的数据也被归并为“个人数据”，比如生物识别数据——指纹、虹膜、DNA数据等；再比如IP地址码, MAC地址码，Cookie信息等，这些信息以往被认为是网络设备信息或网络行为信息，GDPR将其归类到“个人数据”。《个人信息保护法（草案）》的“个人信息”，虽然与GDPR的“个人数据”叫法不同，但实际上概念趋向一致，界定标准也几乎完全类似—— “个人信息是以电子或者其他方式记录的与已识别或者可

识别的自然人有关的各种信息”，同样采取“识别说”为基础，拓宽了个人信息的范畴。企业为了满足合规，必须拥有强大的敏感数据识别能力，能发现各种个人相关的信息以及敏感数据子类别，同样具有分类能力，比如对个人信息主体按照国家归属地进行分类，按照不同儿童和成年人的年龄范围进行分类，以及敏感度分类等。

2) 个人数据/个人信息保护的技术措施

GDPR明确指出保护过程可采取加密或假名化两种措施：加密可保障数据存储和传输过程的安全性，降低数据被非法窃取和泄露的风险；而假名化是GDPR推荐一种“无损的”数据脱敏方式，它对个人数据的标识符信息（比如姓名、身份证号）通过哈希等手段重新命名，同时将真实的标识符-“重命名”映射表与假名化后的个人数据分开存储，以降低隐私泄露风险，同时保证个人数据的完整性。《个人信息保护法（草案）》明确指出可应用加密或去标识化安全技术措施，其中去标识化相比GDPR假名化更为宽泛，去标识化在企业通常称为“数据脱敏”，不仅包括假名化、还包括数据屏蔽、数据泛化、量化、置换等处理方式。这些意味着企业在存储、处理这些个人数据，需采取数据层面的保护措施进行安全防护。

3) 数据权利请求与响应机制

GDPR赋予用户个人数据的知情权、访问权、修改权、遗忘权等各项数据权利，相应地，企业必须响应用户的数据权利请求，比如用户行使“遗忘权”时，企业必须提供删除数据的界面与入口，并执行相关处理操作与流程，以及对用户输出响应报告。且GDPR明确规定企业处理一般请求的响应时间是一个月，复杂请求的响应时间可延长至两个月。《个人信息保护法（草案）》首次全面赋予个人信息主体各项数据权利，包括知情权、决定权、查询权、更正权、删除权等，同时明确指出企业应当建立个人行使权力的申请受理和处理机制。对于响应时间，该草案未明确指出，但《个人信息安全规范》（GB/T 35273-2020）提出响应的时间是30天内（差不多是1个月）。这些促使企业必须建立个人信息请求运营机制，并需要使用流程自动化处理方式。

2 合规驱动下的数据安全技术盘点

Gartner今年7月份将数据安全（Data Security）与隐私（Privacy）作为安全的两个细分领域，分别发布了2020年数据安全成熟度曲线[5]、2020年隐私成熟度曲线[6]，后者与隐私合规性紧密相关。实际上，隐私包含数据安全领域大部分的技术栈，同时也包含新型技术，比如主体权利请求（Subject Rights Request, SRR）、同意与偏好管理（Consent and Preference Management, CPM）等（一般地，国内习惯将隐私并入到数据安全的范畴，将相关技术都统称数据安全技术，本文沿用这种叫法）。

Gartner发布的2020年隐私成熟度曲线，涵盖了35种数据安全相关技术，种类丰富且繁杂，分别处在创新触发期、期望顶峰期、幻想破灭期、稳步爬升期和生产成熟期五个阶段。其中超过70%技术处在稳步爬升期，说明该领域创新技术活跃，有巨大的发展空间，具体如表1所示。

从作用和应用场景角度看，笔者认为35种数据安全技术可分为五大类：

1) 数据安全治理相关

包含多种数据技术组合，以及融合非技术的组织管理措施。比如数据安

全治理（Data Security Governance，DSG）、隐私影响评估（Privacy Impact Assessment, PIA）、数据泄露响应、数字道德、隐私设计（Privacy by design, PbD）和IT风险管理方案。

2) 敏感数据全生命周期的安全防护

包括数据分类、文件分析（针对非结构化敏感数据的识别）、动态脱敏（DDM）、保留格式加密（FPE）和数据销毁（Data sanitization）。

3) 用户隐私权响应与评估合规

包括主体权利请求（SRR）、同意与偏好管理（CPM），可以自动化处理和响应用户提出的数据访问权和删除权等各项权利，以及隐私设计（Privacy by design, PbD），用于在产品设计时考虑隐私合规与可用性问题等。

4) 隐私增强计算类技术

包括差分隐私（DP）、安全多方计算（SMPC）、同态加密（HE）、零知识证明和机密计算（包括TEE）等技术。

5) 其他

包括重点领域的数据安全技术，比如移动终端威胁防御、云环境、5G、区块链的敏感数据保护。

表1 Gartner 2020年隐私成熟度曲线涵盖的相关技术

技术成熟度	数据安全相关技术
创新触发期 (Innovation Trigger)	机密计算、数据安全治理（DSG）、同态加密（HE）、差分隐私（DP）、主体权利请求（SRR）、零知识证明（ZKP）、5G 安全、合成数据、区块链的数据安全
期望顶峰期 (Peak of Inflated Expectations)	数据泄露响应、安全多方计算（SMPC）、同意与偏好管理（CPM）、去中心化实体、数字道德、文件分析、隐私影响评估（PIA）、数据分类
幻想破灭期 (Trough of Disillusionment)	保留格式加密（FPE）、人格化、隐私设计（PbD）、PHI 个人医疗隐私同意管理、移动终端威胁防御、云数据保护网关、隐私管理工具
稳步爬升期 (Slope of Enlightenment)	数据销毁（Data sanitization）、安全即时通讯、电子取证软件、IT 风险管理方案、云访问安全代理（CASB）、动态脱敏（DDM）、云应用程序发现
生产成熟期 (Plateau of Productivity)	数据库审计与防护（DAP）、云安全评估、数据库加密

3 合规视角下的数据安全发展趋势观察

在隐私法规的强有力推动下，国内外数据安全相关技术和产品得到快速发展，逐步形成以“合规遵循”为主的安全细分领域。据2019年11月Gartner的一份预测报告指出，预测在2023年之前全球80%以上的企业将面临至少一项以隐私为重点的数据安全保护规定，并且在合规上的投入将突破80亿美元。由此可见，数据安全合规未来仍然有广阔的市场应用前景。下面对前文提到的数据技术的发展趋势分别进行分析。

观察1：欧美GDPR /CCPA驱动，用户数据权利响应自动化等相关技术发展迅速

全球一些隐私法规赋予数据主体（用户）自由访问、修改和删除个人数据等权利，相应地，要求企业必须在规定的时间内对用户提出的请求进行处理和响应，比如GDPR要求的时间一般为1个月，而CCPA是45天。快速响应数据主体权利请求（Subject Rights Request, SRR）对多数企业是一项极大的挑战。据调查，约有三分之二组织人工处理单个SRR需要两周以上的时间，且平均消耗成本高达1400美元。那么，在合法时间内响应高并发的SRR，传统手工操作是一项困难任务。RSAC 2020创新沙盒比赛中，Securiti.AI一举夺得冠军，它主推自动化的SRR、CPM等用户数据权利响应类产品；另外RSAC2018的创新沙盒的冠军——BigID，它同样聚焦在该类隐私合规产品中；另一家非常著名的创业公司OneTrust有一块很大的业务也是隐私合规性产品，与Securiti.AI几乎重合。这三家初创安全公司融资累计规模超过6000万美元。可以侧面可以反映出，用户数据权利响应产品在国外十分火热，已经发展成为一块稳定的安全市场。

这些产品主要使用了流程自动化以及多种人工智能技术：其中流程自动化可帮助企业的数据安全运营团队从繁琐重复的手工处理“请求-响应”升级为程序的自动化处理，一方面可降低运营成本，另一方面降低由于响应时间延误带来的违规风险；而人工智能技术方面，使用自然语言处理技术（NLP）识别非结构化的敏感数据，使用知识图谱技术关联数据主体所有相关信息，同时使用对话机器人技术方便自动化处理一些提问需求。具体参考《Securiti.ai—解决隐私合规痛点的一站式自动化方案》。

我国《个人信息保护法（草案）》赋予个人包括知情权、决定权、查询权、更正权、删除权等，同时指出“个人信息处理者应当建立个人行使权利的申请受理和处理机制”，但尚未规定具体的时间，而在国标《个人信息安全规范》（GB/T 35273-2020）提出响应的时间是30天内。随着法规的完善，可预计国内SRR、CPM隐私合规技术与市场正逐步形成。

代表公司：Securiti.ai、BigID、OneTrust

观察2：合规基础产品——敏感数据识别、数据脱敏市场日趋成熟

无论是欧盟GDPR、美国CCPA，还是我国的《个人信息保护法（草案）》，均明确表示保护的数据对象是个人数据（或称为个人信息），企业必须履行该类数据的安全保护义务。为了遵循合规，企业第一步是需要识别出存储和流动的各类敏感数据，不仅包括个人基本信息，包括用户姓名、身份证号、手机号等信息，还包括一些个人敏感数据，比如医疗隐私、金融隐私和网络行为的隐私（比如Cookie信息）等。这些敏感数据第一步需要识别。目前已经发展多种敏感数据识别方法：① 基于正则的识别；② 基于关键词库的识别；③ 基于数据相似度的识别；④ 基于机器学习的识别。目前前两种方式在工业界发展较为成熟，一般建立相对全面的规则库或字典。后两种方式通常应用前两种无法解决的敏感数据场景，比如很难直接定义规则或关

键词。第③方法首先从参考数据提取一些特征，然后将其他数据使用同样处理方法后，进行相似度比较，超过一定阈值当作同一类数据；第④方法利用机器学习的强大学习与预测能力，收集足够的样本并进行类别标注，进行模型训练，完成后部署模型自动化识别新数据的类别。识别完成后，为降低敏感数据在二次使用和流通过程（非生产环境，比如数据分析、测试等）的法规风险，大量的数据脱敏需求应运而生。数据脱敏按处理结果是否可还原可分为可逆脱敏和不可逆脱敏技术。可逆脱敏可以理解为企业通过建立一些敏感词的映射表替换为其他非敏感数据，通过反向映射表可将脱敏数据恢复为原始数据。不可逆脱敏技术包含的策略丰富灵活，包括取整、量化、泛化、屏蔽、截断、散列和加噪等。按照使用场景，可将脱敏分为静态脱敏 (Static Data Masking, SDM)、动态脱敏 (Dynamic Data Masking, DDM)。静态脱敏一般用于非生产环境中（测试、统计分析等），动态脱敏一般用于生产环境中。目前静态脱敏技术已经发展较为成熟，而动态脱敏近年来也相关产品落地。

作为两类基础性的合规产品——敏感数据识别和数据脱敏，国内外市场日趋成熟。国内外多家安全厂商在此有所布局，如大型IT公司Microsoft、IBM推出了敏感数据识别和数据脱敏产品，初创公司Securiti.ai、BigID推出了大规模敏感数据的识别产品，并通过AI驱动实现半自动化或自动化扫描和发现。国内绿盟科技推出了IDR产品，可应用在传统数据库和大数据平台的敏感数据发现与分类分级场景中，安华金和推出数据库脱敏相关产品，可应用结构化数据的脱敏应用中。

代表公司：Microsoft、IBM、Securiti.ai、BigID、安华金和、绿盟科技

观察3：合规与数据利用业务场景紧密结合，隐私增强计算技术与应用不断涌现

大数据时代，敏感数据的高频使用和流通，数据既要安全也要求业务利用，这给传统以加密为核心的数据安全技术带来了巨大的挑战。为了满足合规和数据利用的双重需求，促进一批与业务场景紧密结合的新型数据安全技术的产生和发展，包括同态加密、安全多方计算、联邦学习、差分隐私等。由于这些技术不仅可保证原始数据不被泄露（不可见），而且在具体某些业务场景（如聚合、集合运算以及AI建模）保证数据的可用性，工业界习惯将它们形象称为“可用不可见”技术。Gartner将这些技术统称为隐私增强计算（Privacy Enhanced Computation）技术，并将其与随处运营、人工智能工程化等作为 2021年六大重要战略科技趋势。国内外均在此领域有布局：Google的联邦学习及在Android端

应用；Apple在iPhone手机的数据采集中使用了本地化差分隐私技术；RSAC 2018创新沙盒亚军——Duality公司，在定制服务器实现商业化的同态加密方案；阿里主打安全多方计算技术以及平台；百度、腾讯和微众银行等分别推出联邦学习框架并应用在了隐私数据联合建模场景。

代表公司：Google、Apple、Duality、阿里、腾讯、百度、微众银行

观察4：数据安全治理框架与技术方案百家争鸣

传统一两种数据安全技术和措施，无法解决应对内部和外部数据安全威胁，以及合规性和业务带来的挑战。为了应对挑战，Gartner在2017数据安全与风险管理峰会上提出安全治理（DSG）的概念与方法论。数据安全治理——以“数据安全”为核心的综合治理体系，它涉及法规、场景、技术、产品、组织管理以及各类标准流程、策略配置等。微软也提出了针对隐私、保密和合规性的数据治理框架 (Data Governance for Privacy Confidentiality and Compliance, DGPC)，分别从人员、流程和技术这三个角度出发。IBM提出的数据安全和隐私解决方案采用敏感数据发现与分类、评估漏洞、监控与审计等分层方法实现数据安全性。在国内，多家企业提出各自的数据安全治理方法论或数据安全解决方案。比如，阿里提出了DSMM模型，它以数据为中心，数据生命周期为主线，针对数据生命周期各阶段建立全面的数据保护，并对能力成熟度进行定级；安华金和提出了数据安全治理通用框架，框架从数据安全治理机制、数据安全生命周期管理、数据安全技术部署开展数据安全治理与建设；绿盟在Gartner数据治理框架基础上，结合客户的数据安全防护需求，对实际情况进行研究和实践，也建立一套完整科学的方法体系——数据安全解决方案。该体系分为五个基本治理步骤——“知”、“识”、“控”、“察”、“行”，五个步骤分别采用不同的数据安全技术与管理措施具体参考《拨开云雾见天日——数据安全治理体系》。

代表公司：Microsoft、IBM、阿里、安华金和、绿盟科技

4 小结

在全球相关法规的推动下，如欧盟GDPR，美国CCPA，以及我国最近发布的《个人信息保护法（草案）》，隐私合规逐步成为企业数据安全建设与治理重要驱动力。在法规监管不断强化的背景下，企业必须主动进行合规性建设，结合自身业务场景与风险，实施体系化的数据治理与建设，在数据的全生命周期结合安全需求实施一项或者多项技术与措施以应对数据安全风险。在一些新的数据安全场景，尤其是数敏感数据的安全共享计算，该领域创新技术不断，包括安全多方计算、联邦学习、差分隐私，唯有通过跟踪和探索这些新技术的发展，才能更好应对新场景中带来的新的数据安全问题、新的安全风险以及合规性挑战。

参考资料

1. General Data Protection Regulation (GDPR), https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv:OJ.L_.2016.119.01.0001.01.ENG&toc=OJ:L:2016:119:TOC.

2.《个人信息保护法（草案）》，http://www.npc.gov.cn/

3. 金融信息安全成监管重点 央行开千万元级罚单护航， https://finance.sina.com.cn/stock/jhzx/2020-10-30/doc-iiznezxr8873053.shtml

4. 绿盟科技，《数据安全白皮书2.0》

5. Gartner, Hype Cycle for Privacy, 2020

6. Gartner, Hype Cycle for Data Security, 2020

数据安全治理体系之解析

IDC的“大数据摩尔定律”表明，人类社会活动产生的数据一直在以每年50%的速度增长。也就是说，全球数据量将在每两年翻一番。据IDC统计与研究，全球数据量已经进入了ZB (1ZB=〖10〗^12 GB)级别[1]。随着企业业务发展和扩大，应用环境的数据越来越庞大，多种多样、复杂多变。面临的数据安全问题和威胁越来越突出和严峻，不仅有来自外界的攻击，也有内部管理或错误配置等引发的数据窃取或敏感信息泄露。

图1 当前复杂数据应用环境：正如浩瀚如烟、杂乱无章的图书[2]

在当前错综复杂的海量数据环境下，如何更好地开展数据安全建设和防护？有条不紊地做好法律合规工作？更好地应对数据安全的内部和外部等威胁与

挑战？基于此，获得一本“武功秘籍”（掌握一套科学的数据安全实践体系）对于企业来说是十分重要且必要。

图2 数据安全领域的“武功秘籍”：数据安全治理体系

本文是“大数据时代下的数据安全”系列的最后一篇：实践体系篇。国内外多数安全公司在数据安全领域都有一本属于自己的“武功秘籍”，本篇将重点介绍比较有代表性的三本：Gartner数据安全治理框架、绿盟数据安全解决方案，以及数据安全能力成熟度模型。

1 Gartner数据安全治理框架

数据安全治理 (Data Security Governance，DSG)最早由Gartner在 2017安全与风险管理峰会上提出。数据安全治理从字面理解：“数据安全”+“治理”。“治理”不是“管理”，从范畴来说，前者来说更大，强调以“数据安全”为核心的系统性/综合性的过程。因此，“数据安全治理”可以简单理解为针对数据安全的综合治理过程与体系。

Gartner 认为数据安全治理是从决策层到技术层，从管理制度到工具支撑，自上而下贯穿整个组织架构的完整链条。组织内的各个层级之间需要对数据安全治理的目标达成共识，确保采取合理和适当的措施，以最有效的方式保护数字资产[3-5]。其安全治理框架如下图所示，“由上而下”，一共分为5步：第一步是治理工作的开始，需要进行总体上层的设计，评估数据安全的业务风险：风险时什么？在哪里？为什么？以及如何？ 具体来说，考虑和平衡经营策略、治理、合规、IT策略和风险容忍度之间的关系；第二步对数据进行分类和分级，对不同类别的数据类型以及敏感度进行打标签；第三步根据前一步的数据标签，结合不同用户类别，对不同的数据类型、敏感度制定相应的策略，实施更精细粒度的数据管理；第四步根据前几步分析的场景需求，确定需要的数据安全产品，比如加密（Crypto）、数据审计（DCAP）、数据防泄漏系统（DLP）、云防护代理（CASB）、身份权限管理和访问控制（IAM）以及用户行为分析（UEBA）。最后一步根据前面分析的策略，为所有产品编排策略，包括数据库、大数据系统、文档、云端和终端等环境的数据安全管理和防护产品。

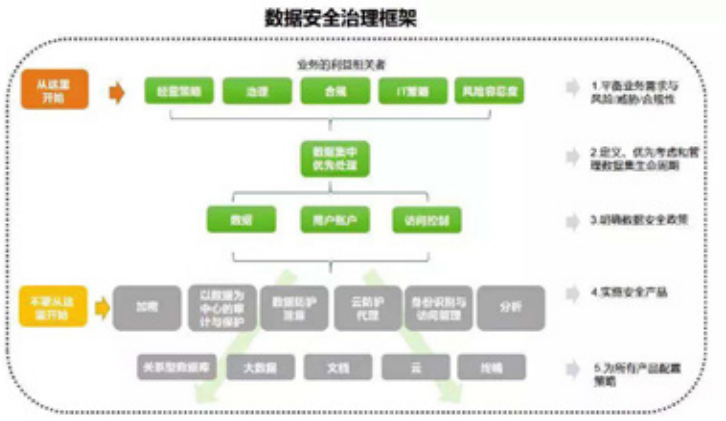


图3 Gartner数据安全治理框架[6]

2 绿盟数据安全解决方案

绿盟数据安全解决方案在Gartner数据治理框架基础上，结合客户的数据安全防护需求，对实际情况进行研究和实践，也建立一套完整科学的数据安全治理方法体系[6-8]。该体系分为四个基本治理步骤——“知”、“识”、“控”、“察”。下面对这四个“动词”分别详细解析：

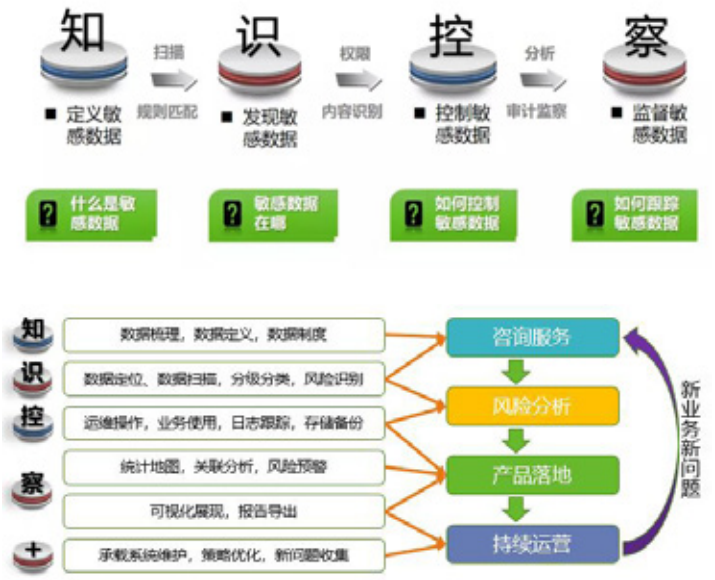


图4 绿盟数据安全解决方案[6-7]

知：分析政策法规，如中国的《中华人民共和国网络安全法》、《电信和互联网用户个人信息保护规定》、《个人信息安全规范》、欧盟的《General Data Protection Regulation》等；及时梳理业务及人员对数据的使用规范，以及定义敏感数据，包括类别、敏感度等；

识：根据定义好的敏感数据，利用工具对全网进行敏感数据扫描发现，对发现的数据进行数据定位、数据分类、数据分级。这一步十分重要，直接决定后续治理步骤和数据保护的质量；

控：根据敏感数据的级别，设定数据在全生命周期中的可用范围，利用规范和工具对数据进行细粒度的权限管控；

察：对数据进行监督监察，保障数据在可控范围内正常使用的同时，也对非法的数据行为进行了记录，为事后取证留下了清晰准确的日志信息。如部署数据

审计（DCAP）、数据防泄漏系统（DLP）、以及用户行为分析（UEBA）等安全产品。

3 数据安全能力成熟度模型

数据安全能力成熟度模型(Data security capability maturity model, DSMM)最早由阿里提出，目前已经完成标准化（《信息安全技术 数据安全能力成熟度模型》（GB/T 37988-2019）），2020-03-01即将实施。在该项标准中，DSMM由三方面构成，其架构关系如图5所示 [8]：

(1) **数据生命周期安全：**围绕数据生命周期，提炼出大数据环境下，以数据为中心， 针对数据生命周期各阶段建立的相关数据安全过程域体系。

(2) **安全能力维度：**明确组织机构在各数据安全领域所需要具备的能力维度， 明确为制度流程、人员能力、组织建设和技术工具四个关键能力的维度。

(3) **能力成熟度等级：**基于统一的分级标准， 细化组织机构在各数据安全过程域的 5 个级别的能力成熟度分级要求。

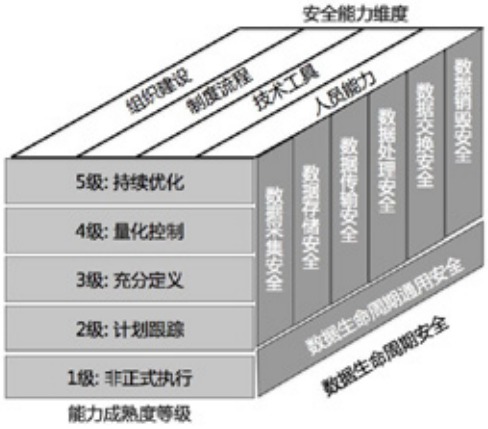


图19 数据安全能力成熟度模型架构[8]

以数据生命周期为主线，其数据安全过程域体系，分为数据生命周期通用的安全和各生命周期阶段下的安全，包含一系列相应的内容、方法和策略，如图11所示[8]。

图20 数据安全过程域体系[8]

总体来说，数据安全能力成熟度模型以数据为中心，拓展了一切与数据安全相关的维度，涵盖的内容十分丰富且流程十分完善。但这意味着具体实践与落地需要投入更多的资金、实施更长的时间等成本满足该项标准，这是一项挑战。

4 小结

在大数据时代，面对庞大、多种多样、各类数据相互掺杂等的复杂数据环境，传统的一两种数据安全技术/方案，难以应对内部或外部原因引起各种各样新出现的威胁与挑战。掌握一套科学且系统的数据安全实践体系，“以不变应万变”——这对于一个高度数据化的企业来说是十分重要且必要。像数据治理一样，数据治理正被越来越多安全企业提起、诠释以及完善。虽然不同企业的方案和流程各有差异，但笔者认为，数据安全治理是体系化的指导思想，是解决现有的、未来的数据安全问题一般方法论的抽象，这些正是它们的共同目的。数据安全治理的目的决定它是一个综合/系统的过程，涉及法规、场景、技术、产品、人员管理以及各类标准流程、策略配置等。

虽然当前我国数据安全相关法规还在建立和完善中、惩罚机制尚且没有具体给出量化措施。但笔者认为，数据安全的建设应该是一个主动过程，而不是一个被动过程：一方面主动的建设比被动建设避免一些不必要的安全风险以及经济损失；另一方面，数据安全治理需要一个漫长的周期，若等到法规的完善和强制执行时候才开始，却为时已晚。因此，主动建设和投入十分重要。

数据分类与分级是数据安全治理的开始，也是其关键的第一步。通过定义、扫描、测绘、梳理、分类和分级等，能准确地掌握敏感数据在哪里？风险在哪里？对应风险级别？了解到企业敏感数据分布的概貌，从而更好进行治理。绿盟为了帮助客户更好地完成第一步，在RSA2019年上已经发布了大数据安全产品——绿盟敏感数据发现与风险评估系统（绿盟IDR：NSFOCUS Insight for Discovery and Risk），具有智能的数据分类分级、全网的数据资产测绘、实时的数据流转测绘和全面的数据安全风险评估等功能[9]。欢迎咨询，欢迎了解（《新品发布·绿盟科技IDR敏感数据发现与风险评估系统》）。

参考资料

1. Gantz J, Reinsel D. The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east[J]. IDC iView: IDC Analyze the future, 2012, 2007(2012): 1-16.

2. Gartner Summit 2018: Data Classification

3. Gartner Summit 2018: State of Data Security 2018

4. Gartner Summit 2019: Outlook for Data Security 2019

5. 数据安全治理委员会，数据安全治理白皮书，2018

十种前沿数据安全技术，聚焦企业合规痛点

2020年7月和10月，我国陆续发布两部重磅级的法规草案——《数据安全法（草案）》和《个人信息保护法（草案）》。欧盟于 2018 年实施《通用数据保护条例》（GDPR），美国于 2020 年实施《加州消费者隐私法案》（CCPA），日本于2020年6月通过修订版《个人信息保护法》。随着全球数据安全法规监管的不断强化，合规性问题不得不纳入企业数据安全建设考虑范围。然而，法规对企业更高的安全要求，这给传统的数据安全防护技术与措施带来了前所未有的挑战。

在此背景下，绿盟科技近日发布《拥抱合规、超越合规：数据安全前沿技术研究报告》。在报告中，选取业界最为前沿与创新的十种数据安全技术，对其技术原理与应用进行全面的梳理与分析，包括处于学术前沿的差分隐私、同态加密、数据匿名；行业内炙手可热的安全多方计算、联邦学习等。这些新兴技术，为企业的数据安全建设带来新的思路与方案——助力其在满足业务需求的同时解决合规的痛点与难点。

1 简介

数据安全建设离不开具体的业务场景，数据安全技术需要从应用场景出发。根据企业的业务系统与应用、以及数据分布范围的不同，我们将数据安全建设分为三类场景：

1) 用户隐私数据安全合规；

2) 企业内部数据安全治理；

3) 企业间数据共享与计算。

如图1所示，上述三大类场景根据具体业务与功能的不同，可进一步细分一些子场景。各个子场景不仅有自身内部安全需求，也有相应的合规性要求，具体可对应到欧盟GDPR条款，以及我国已实施的《网络安全法》的数据安全相关条款。后续三个章节将从三类场景以及子场景的应用需求与合规挑战出发，研究与分析如何基于前沿技术，实现超越合规，解决安全痛点。

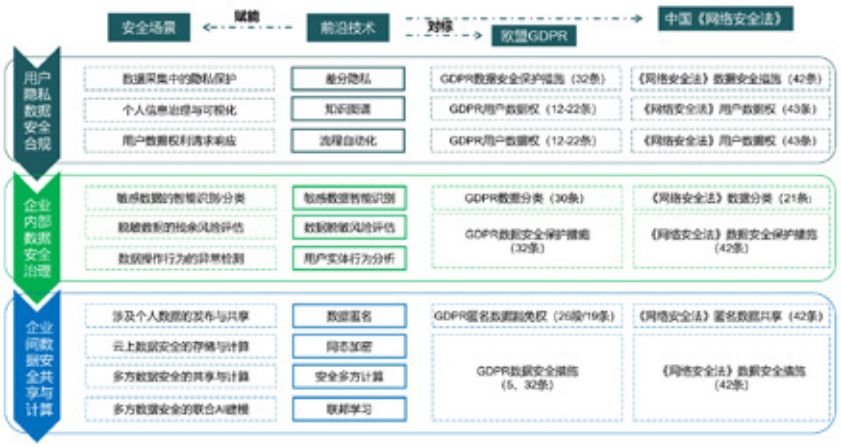


图1 超越合规：数据安全场景-前沿技术图谱

2 前沿技术+用户隐私数据安全合规

在该类场景中，企业需解决用户隐私数据的采集、以及数据权利请求响应的合规性问题，可引入下述创新技术：

1) 差分隐私

技术原理：差分隐私是一种基于噪声机制的隐私保护技术。在本地差分隐私模式下，每一个用户终端都会运行一个差分隐私算法，每一个终端采集的数据都会加入噪声，然后将其上传给服务器；服务器虽然无法获得某一个用户的精确数据，但通过聚合与转换可以挖掘出用户群体的行为趋势。

合规遵循：GDPR的 32条和《网络安全法》的42条。

行业应用（代表公司）：Google、Apple，其中Apple通过差分隐私可挖掘到iPhone用户使用表情的频率分布，但无法获得具体某一个用户的确切隐私。

图2 iPhone差分隐私技术应用[1]

2) 知识图谱

技术原理：知识图谱最早用于搜索引擎和社交网络，它简单可以看成是一种基于图的数据结构，由节点和边组成，每个节点是一个实体，每条边是两条实体之间的关系。由于个人数据治理关键是个人数据实体识别，以及相关属性与处理流程的关联，引入知识图谱技术成为必然。通过知识图谱技术，可帮助企业了解所在敏感数据的位置，是如何被使用的，以及它的合同、法律和监管义务，达到个人信息治理与可视化作用。

合规遵循：可满足GDPR的12-22条和《网络安全法》的43条。

行业应用（代表公司）：RSA 2020创新沙盒冠军Securit.ai公司，基于知识图谱技术实现了个人数据图谱应用。

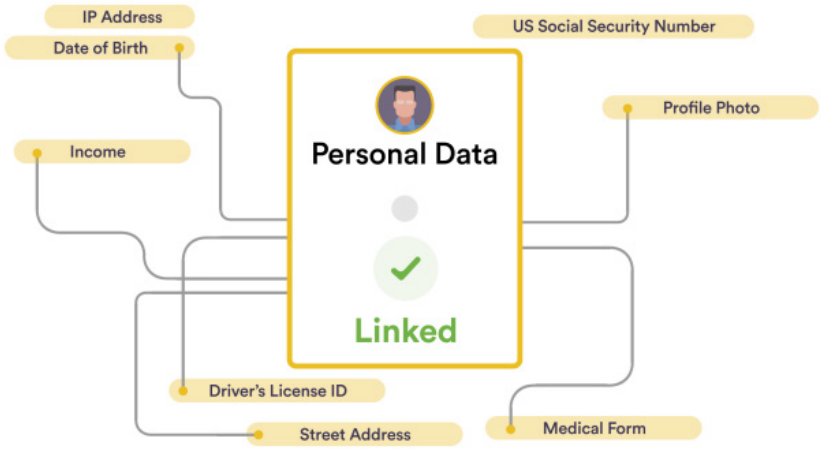


图3 Securiti.ai的个人数据图谱应用[2]

3) 流程自动化

技术原理：用户数据权利请求响应是欧美等国外企业重要的隐私合规检查项。流程自动化技术可帮助企业的数据安全运营团队从繁琐重复的手工处理“请求-响应”转为自动化处理，一方面可降低人工的运营成本，另一方面可减少由于响应时间延误（GDPR规定一般为一个月）带来的违规风险。

合规遵循：GDPR的12-22条和《网络安全法》的43条。

行业应用（代表公司）：Securit.ai、BigID和OneTrust等。

3 前沿技术+企业内部数据安全治理

在该类场景中，企业需解决内部敏感数据治理的安全与合规问题，可引入下述技术解决安全与合规问题，可引入下述创新技术：

1) 智能敏感数据识别

技术原理：传统基于关键词、正则匹配的敏感数据识别方法不够智能，易出现漏检（尤其是在文档等数据）。引入相似度计算、聚类、监督学习等智能方法，提升识别能力与检测效果。

合规遵循：GDPR的30条和《网络安全法》的21条。

行业应用（代表公司）：Securit.ai、BigID等。

2) 数据脱敏风险评估

技术原理：数据脱敏在企业进行广泛应用，然而不同脱敏方法的安全效果不同。通过对脱敏数据集的身份标识度和隐私泄露风险进行定量地评估与刻画，实现风险管理和控制。

合规遵循：GDPR的32条和《网络安全法》的42条。

行业应用（代表公司）：Privacy Analytics、绿盟科技等。

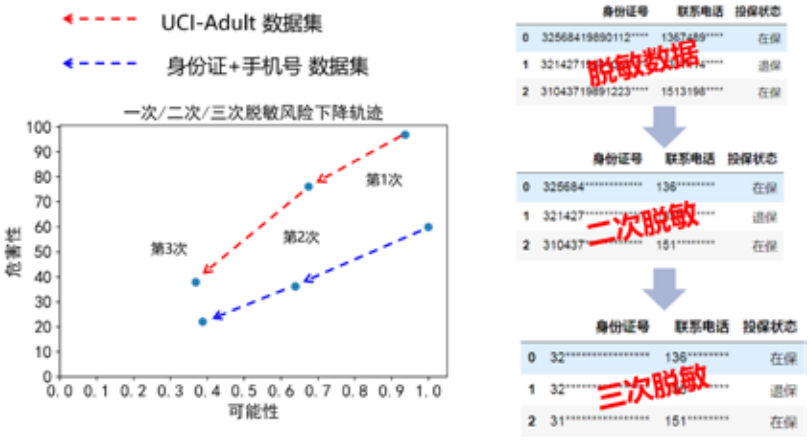


图4 绿盟科技的数据脱敏风险评估应用

3) 用户实体行为分析

技术原理：通过对用户实体持续的画像与建模，并建立正常用户行为基线，

从海量收集的安全数据中发现数据泄露等异常行为。

合规遵循：GDPR的32条和《网络安全法》的42条。

行业应用（代表公司）：Splunk、绿盟科技等。

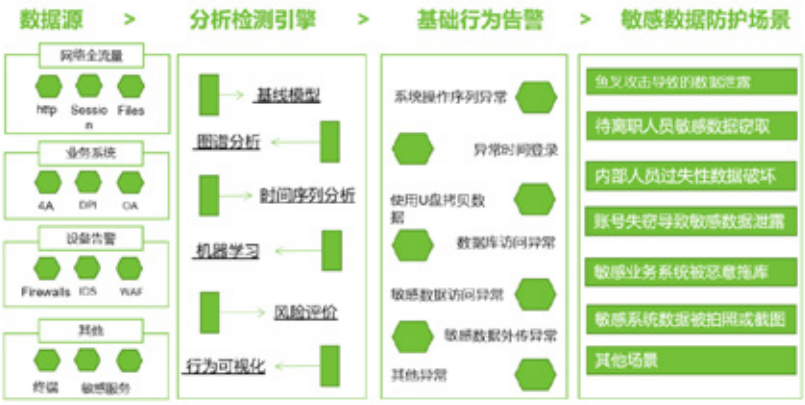


图5 绿盟科技的UEBA数据安全防护方案

4 前沿技术+企业间数据共享与计算

在该类场景中，企业需解决企业之间的数据安全共享与计算的安全与合规问题，可引入下述创新技术：

1) 数据匿名

技术原理：对个人信息进行泛化和屏蔽等处理，使得对应的个人信息主体无法被识别，以达到“匿名”的效果，包括K-匿名、L-多样性和T-近似性等技术。

合规遵循：GDPR的前言26段和19条，以及《网络安全法》的42条。

行业应用（代表公司）：Immuta、Privitar、 Anonos、绿盟科技等。

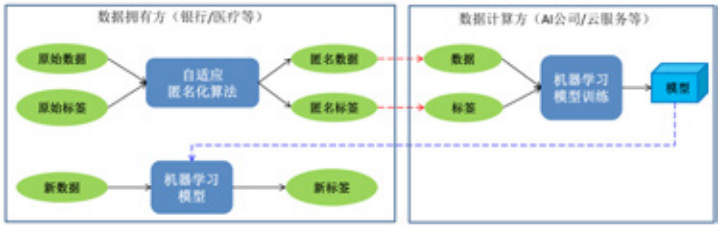


图6 绿盟科技的自适应匿名化算法应用

2) 同态加密

技术原理：明文数据经过同态加密后得到的密文数据，在不解密情况下仍然可执行密文数据的处理与操作。敏感数据在同态加密与计算环节处于加密状态，实现了数据的计算，同时保障了安全性。

合规遵循：GDPR的前言5条和32条，以及《网络安全法》的42条。

行业应用（代表公司）：Duality 等。

图7 Duality的同态加密平台在金融数据共享应用（图引自[3]）

3) 安全多方计算

技术原理：在参与方互不信任的情况下进行协同计算，在保证计算结果正确性同时不泄露任何一方输入的原始数据和状态数据。

合规遵循：GDPR的前言5条和32条，以及《网络安全法》的42条。

行业应用（代表公司）：Google、蚂蚁金服等。

4) 联邦学习

技术原理：多个参与方（如企业、用户移动设备）在不交换原始数据情况下，即在隐私保护前提下，实现联合机器学习的建模、训练和模型部署。

合规遵循：GDPR的前言5条和32条，以及《网络安全法》的42条。

行业应用（代表公司）：Google、Apple和微众银行等。

5 小结

随着全球数据隐私法规的密集发布，包括欧盟GDPR，美国CCPA，国内的《网络安全法》，以及今年发布的《数据安全法（草案）》、《个人信息保护法（草案）》，合规性成为了企业数据安全建设与治理的重要驱动力。在合规视角下，数据安全的内涵在合规与业务安全双重需求驱动下不断外延和扩展，数据安全的覆盖的应用场景将变得更加多样化，给传统的数据安全技术与方案带来了巨大的挑战。如何实现破局？本文简要介绍的十种新兴的数据安全技术，可为破局新场景新挑战带来一些思路与启发——助力企业在满足安全合规同时创造更大的数据价值。

参考文献

[1] Differential Privacy, https://www.apple.com/privacy/docs/Differential_Privacy_Overview.

[2] Securiti.ai homepage. https://Securiti.ai/.

[3] Duality Homepage. https://dualitytech.com/.

【RSA2020 创新沙盒】Securiti.ai—解决隐私合规痛点的一站式自动化方案

1 公司介绍

Securiti.ai成立于2018年11月，总部位于美国加利福尼亚州的硅谷地区。当前融资总额达到8100万美元，最近一次是由General Catalyst领投，Mayfield参与的5000万美元B轮融资[1]。其创始人兼CEO是Rehan Jalil，拥有丰富的网络安全从事经验和管理背景，先后在Elastica和Bluecoat担任CEO，后在赛门铁克云安全部门担任高级副总裁 [2]。公司致力于实现隐私合规的自动化，利用AI和People Data Graph等先进技术，使得隐私合规遵循从传统的繁琐手工操作到一站式的自动化成为可能，以帮助企业快速满足GDPR和CCPA等法律法规 [3]。

2020年1月1日，CCPA（《加州消费者隐私法案》）正式生效，如何快速地且有条不紊地满足该隐私法案，这是众多硅谷巨头（Google、Facebook和Apple等）以及中小企业的一大痛点问题。在此背景下，Securiti.ai凭借应对方案以及技术实力入选2020年RSA大会创新沙盒的前十强。值得一提的是，BigID作为隐私数据治理（遵循GDPR），在2018年的创新沙盒夺得冠军；Securiti.ai作为一家同样主打隐私合规产品（可遵循CCPA）的创新公司，是否能赢得今年RSA创新沙盒的冠军？值得期待！

2 背景介绍

为了保障公民的个人信息与隐私安全，全球掀起了隐私法规的立法热潮。2018 年 5 月 25 日，欧盟正式颁布《通用数据保护条例》（General Data Protection Regulation，GDPR）用以保护欧盟成员国境内企业的个人数据、也包括欧盟境外企业处理欧盟公民的个人数据。受GDPR 影响，全球各个国家推出了

类似的个人信息保护法规，如巴西LGPD、印度PDPB、泰国PDPA等。我国于2017 年 6 月 1 日正式实施《网络安全法》，并于去年发布《数据安全管理办法（征求意见稿）》，对个人信息安全提出诸多规定和约束。2019 年 10 月，美国加州州长正式签署 《加州消费者隐私保护法》（California Consumer Privacy Act，CCPA） 的最终法案，于今年1月1日正式生效[4]。

从GDPR的执法来看，违反的罚款代价是高昂的。例如，法国于2019年1月份罚款 Google公司 5000万欧元，原因是Google 的隐私条款的设计非常难以被用户理解，尤其是在个人化广告推荐上；英国在2019年7月份分别对英航和万豪集团分别开出 1.83 亿英镑和 9900万英镑的罚单，原因均为企业数据防护措施不力导致了数据泄露；德国对网络公司Delivery Hero处以 20 万欧元罚款，原因是客户要求删除个人数据时，却没有及时性应答。对于CCPA的实施，据IAPP和 OneTrust 的CCPA Readiness调查结果显示，74%的受访者认为他们的雇主应该遵循CCPA，但遗憾的是，仅大约2%的受访者认为他们的企业已经完全做好了应对 CCPA 的准备[5]。

这些隐私法规迫使企业对以下一些问题进行思考：存储那些消费者的个人数据？它们分布在那些系统中？是否满足立即响应客户的数据访问权、更新权和删除权等权利？如何解决跨多个应用程序与第三方共享的数据问题？对于多数企业来说，这是一系列的合规性痛点问题。Securiti.ai正抓住这一普遍诉求，并且利用AI和People Data Graph等先进技术，使得繁琐的处理流程变得更加自动化。

3 公司产品与方案

3.1 产品介绍

Securiti.ai的产品称为Privaci，公司CEO Jalil将其描述为“PrivacyOps”解决方案，并亲自写一本书对产品功能和架构思想进行介绍[6]。Jalil 将PrivacyOps概括为它是哲学、实践、跨功能协作、自动化和业务流程的组合，它提高了企业组织可靠且快速地遵守众多全球隐私法规的能力。通俗地说，传统隐私合规性的请求处理是手动的、缓慢的，而PrivacyOps可实现批量地、自动化地处理规性请求，并可定期评估与检测合规性风险，能够帮助企业快速满足全球隐私法规遵循的要求。

公司官网中展示了PrivacyOps的五个子产品：Data Fulfillment Automation、PD Linking Automation、Assessment Automation、 Third Party Risk Assessment和Consent Lifecycle。为了更好地理解，笔者将五种产品归为三类：前两者为应对消费者数据权利（访问权、修改权和更新权等）请求响应的合规处理产品、后两者是合规性风险评估类产品、最后一个为数据授权处理的合规性检查类产品。

3.1.1 Data Fulfillment Automation

CCPA和GDPR等隐私法规赋予数据主体（消费者）数据访问、修改和删除等权利，比如在CCPA的1798.100条款中，规定“企业收到消费者访问个人信息之请求后，应当立即通过邮寄或电子等方式向消费者披露和传送所要求的数据”；GDPR也有类似规定，要求企业必须在一个月内对所有请求进行响应，若请求过于复杂，可延长到两个月。该合规场景可解读为企业需提供两点功能：

① 为消费者提供数据权利请求的窗口；

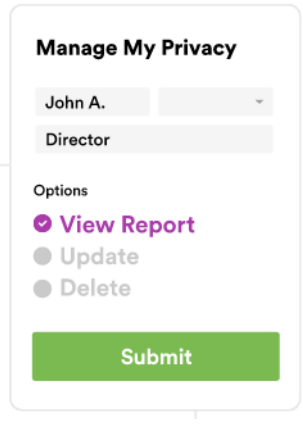
② 收到请求立即或者在规定时间内进行响应。

假设一天有1000个用户请求，若采取手动操作，查询相关系统，并手工制作1000个用户的个人信息数据报告，这个工作量是巨大的，且容易产生操作错误。因此亟需一种流程自动化的方法。

Data Fulfillment Automation（数据权利履行自动化）产品可实现消费者数据权利请求-响应的流程自动化处理，并且可生成合规性审查报告。其产品运行流程如下：

1）构建一个或多个动态请求表单并嵌入到客户的网站中，类似于一个网站插件。通过这个插件，可轻松接受数据主体的请求（Data Subject Request, DSR）。

2）收集DSR请求，并根据用户的身份创建DSR工作流程。为了避免错误处理和欺诈，该过程需验证用户的身份。下图显示John A. 用户希望查看网站收集自己那些数据，除数据访问外，还可对收集的个人数据进行编辑、甚至删除。

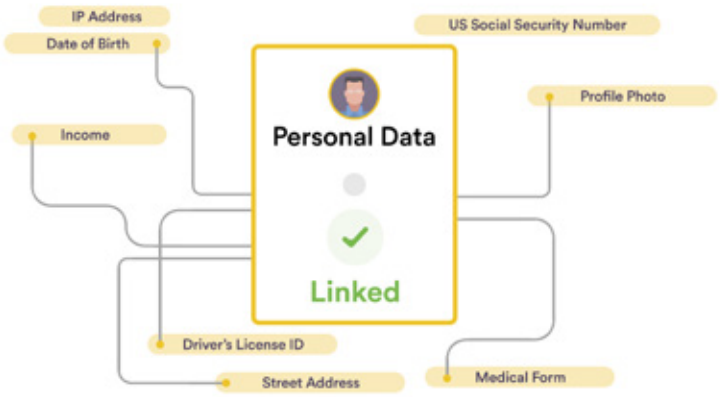


3）DSR工作流程提交给后台，后台有机器人助手Auti，可关联到用户John A.的数据，帮助完成DSR任务并且同步系统。

4）生成DSR履行的审查报告，以证明遵循隐私合规。

3.1.2 PD Linking Automation

PD Linking Automation（个人数据链接自动化）产品通过强大的数据管理能力及People Data Graph技术，可将企业在不同时间、不同系统收集和存储的某个数据主体所有相关数据进行关联，比如数据主体的IP地址、身份证号、驾驶证号、照片、出生年月、住址和收入等个人信息。



根据PrivacyOps book [6]的详细介绍，该关联技术的应用产品主要有以下三种应用场景：

① 辅助前面产品Data Fulfillment Automation，生成个人数据的关联报告。比如，用户向比如Google提出访问个人信息的请求后，Google有多个产品与系统，在浏览器服务器记录用户注册信息和Cookie信息，另外在邮件服务器中也记录了同一用户的个人信息，这两个系统存储的同一个数据主体的信息，但多数企业不会将两个系统进行关联。但GDPR和CCPA要求企业向消费者披露数据主体的相关个人信息，因此关联技术十分重要；

② 跨国企业的个人数据的治理与可视化。跨国企业的数据存储、处理服务器分布全球各地，若将个人信息主体的信息进行关联，并关联到用户的国籍或居住地（欧盟、美国加州），可视化获得数据分布的世界热图，更好地洞察不同国家地区法规的合规性风险；

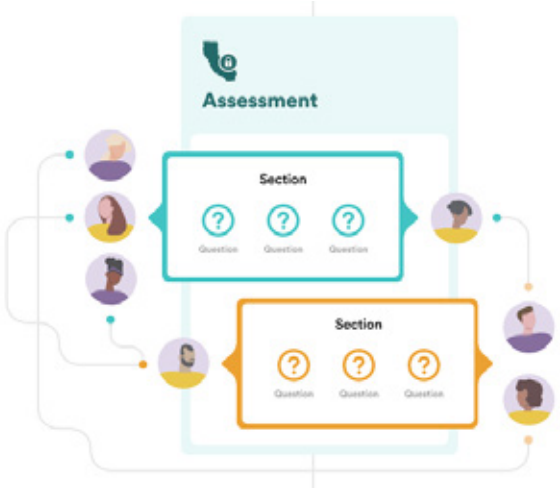
③ 数据泄露后的及时通知受影响的数据主体。例如，对于个人信息数据库（假如无邮件、电话等联系信息）的泄露，通过关联邮件、电话等信息，

可快速通知泄露的用户，满足合规性要求。

通过向聊天机器人Auti提问，可触发操作流程实现自动化执行，并生成宏观的个人数据地图和审查报告。

3.1.3 Assessment Automation

Assessment Automation（内部评估）可为企业内部的系统进行隐私风险评估。产品系统内置不同的合规性模板，如GDPR、CCPA，每一种合规性模板对应各种合规点检查列表和问题。为了高效地完成隐私风险的评估，产品提供了一个协作平台，通过它内部多位安全专家可分工对问题进行检查和回复。协作平台收集所有的输入和检查，最终生成评估报告。一旦报告生成，企业可选择与第三方组织或消费者分享，以证明隐私风险控制能力。



3.1.4 Third Party Risk Assessment

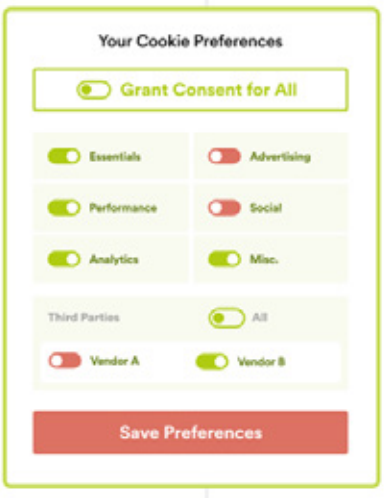
在GDPR中，有两个重要的数据处理组织：数据控制者（Controllers）和数

据处理者（Data Processor），数据控制者是数据主体的第一联系人，负责数据的收集与处理，数据处理者在多数场景下是第三方组织。例如，一家零售机构（数据控制者）采集用户信息，它租赁亚马逊云服务器（数据处理）进行数据的存储和计算。按照GDPR法规，零售机构的法规责任更大，必须慎重挑选挑选数据处理者，以确保它由能力通过适当的技术和组织措施以满足GDPR的要求。CCPA也有类似的规定，当企业与第三方进行个人信息的交互时，第三方发生违发数据行为时，当事企业也承担一定法规责任。特别在大型公司，拥有多个合作商，数据交互和流通越来越大，为了降低法规风险，不仅需保证内部满足隐私合规，也需确保合作的第三方是可信任的，隐私风险控制水平达到安全级别。

Third Party Risk Assessment（第三方组织的风险评估）很好解决以上的一个痛点，它可以同时邀请与企业合作的多个第三方组织，共同接入评估平台进行隐私风险评估，它尽可能利用可获取的数据，包括各个网站的隐私声明（privacy statements），第三方组织安全专家提供合规性证明和检查文件。该产品在生成评估报告同时，也提供了一个统一的隐私风险评分服务。

3.1.5 Consent Lifecycle

Consent Lifecycle（许可生命周期管理）产品可应用与网站的Cookie的数据收集，征求用户的同意，对标多个GDPR、CCPA等法规的数据处理与利用的公开透明原则。首先，提供服务的企业需在自身网站中部署Consent Lifecycle相关插件，如下图所示，它提供了用户授权设置的一个窗口。然后，用户可在设置盘中授权网站的Cookie信息授权应用于那些目的，比如数据分析、广告推荐、社交发现，提供给第三方组织C1企业或C2企业等。那么，如果用户没有授权Cookie信息用于广告，而服务企业却在后台的广告推荐系统利用了该Cookie信息，那么Consent Lifecycle将监控到这一异常行为，在后台触发告警行为，以通知企业停止违反法规的风险行为。

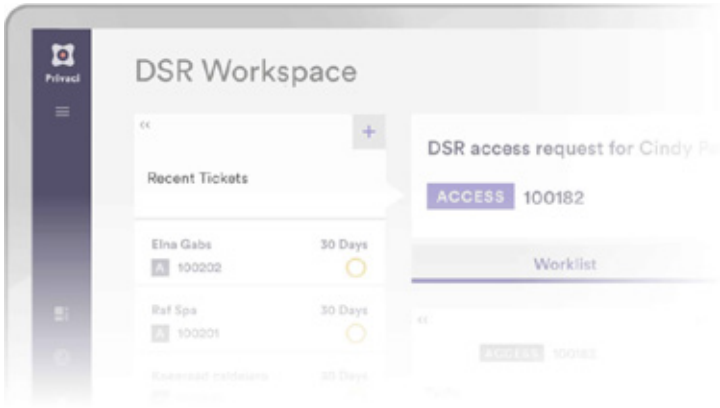


3.2 合规性方案

前文介绍的公司的五个产品，它们进行组合可对标到CCPA（美国加州）、GDPR（欧盟）和LGPD（巴西）隐私合规条款，公司的官网分别提供三种隐私法规遵循的解决方案。下面以CCPA的解决方案为例，对产品功能与对标合规点进行简单介绍，详细合规功能点可访问官网 [3]。

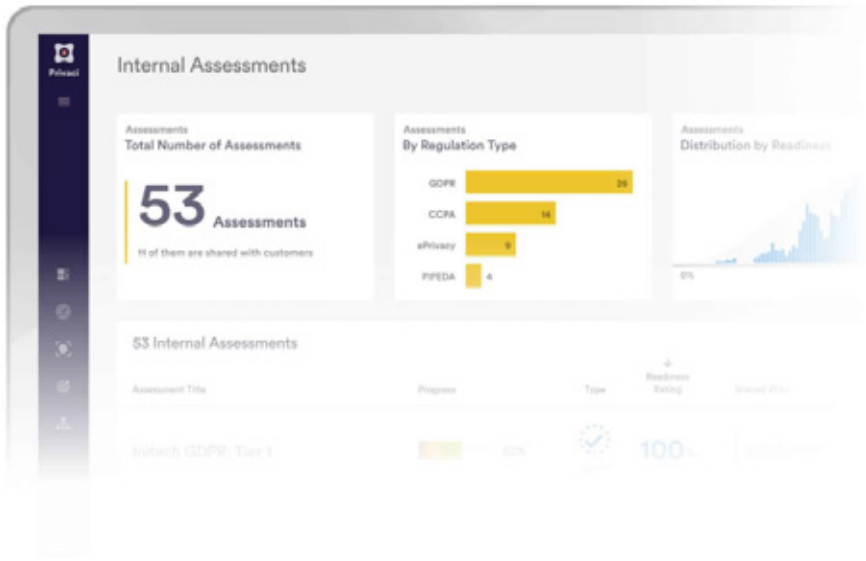
3.2.1 数据主体请求 DSR 自动化处理

CCPA规定消费者具有数据访问、更新、删除的权利。当用户提出合理的权利诉求，PrivacyOps方案可自动化收集、接收以及处理数据主体请求（Data Subject Request, DSR），并自动生成DSR报告。具体对标CCPA的1798.100、1798.105、1798.110、1798.115的数据访问权相关条款。这些规定了企业收到消费者访问个人信息之请求后，应当立即通过邮寄或电子等方式向消费者披露和传送所要求的数据，否则视为违反法规。



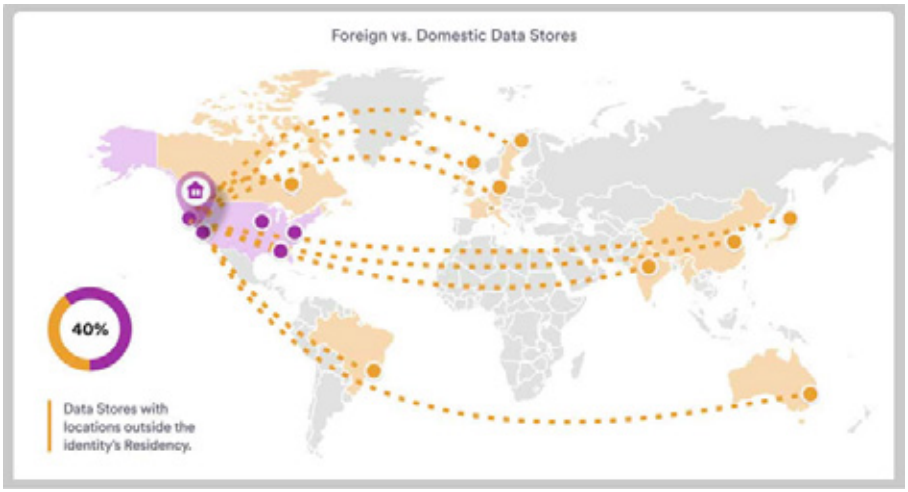
3.2.2 隐私风险评估

通过使用协作的、准备就绪的PrivacyOps评估系统来衡量企业组织的隐私合规性，识别差距并处理风险，以保持符合CCPA监管要求。具体对标CCPA的1798.135.(1)(2)相关条款，要在互联网主页或隐私政策的醒目清晰位置，提供一个命名为“不得出售我的个人信息”，消费者有选择不出售个人信息数据的权利。



3.2.3 个人数据自动链接

发现所有系统中存储的个人信息，并将其链接到个人数据的所有者。通过身份可视化数据蔓延，并基于受试者居住地识别合规风险。CCPA监管的是处理加州居民个人数据的营利性实体，即美国加州外的其他州、甚至国外的跨国企业处理加州居民的用户数据，将受到相关法规风险影响。该功能可视化展示宏观的风险分布位置，对标CCPA的治理和监管要求。



除以上合规要求外，PrivacyOps方案还可满足用户选择不出售数据的权利、默认不选择、检测隐私说明等合规性要求。

4 总结

欧盟GDPR在2018 年正式实施，2019年进入全面执法，多张巨额的企业罚单相继被开出。美国加州隐私法规CCPA在今年1月1月已经生效，与GDPR一样，CCPA同样是一项十分严格的隐私法规，赋予了消费者更多数据权利，比如访问权、修改权、删除权和不出售数据给第三方等，同时提出了企业履行的义务条款。比如，企业收到消费者要求访问个人信息的请求后，应立即采取措施（比如信件或电子方式）向消费者免费披露和提供本节所要求的个人信息（GDPR也有类似的条款）。如何快速地满足遵循CCPA，这是众多硅谷巨头（Google、Facebook和Apple等）以及中小企业面临的一大痛点问题。违反法规意味着罚款与处罚，GDPR罚款的上界是2000万欧元或者4%全球营业总额，而CCPA可具体到每一个消费者，罚款上限为750美元（同时被1万人起诉，那么相当于罚款750万美元）。值得注意的是，GDPR不仅对欧盟，CCPA不仅对美国加州的企业，只要处理欧盟，或者加州的居民的跨国企业，同样将受到法规的域外监管与约束。

Securiti.ai瞄准了这一个普遍的隐私合规性市场与需求，将繁琐的合规性遵循变成了智能的、自动化的处理，可满足企业满足隐私合规的大部分需求。具体来说，Securiti.ai具有以下亮点与优势：

① 融资数额8100万美元是创新沙盒前十强最高的数额 （第二名Obsidian Security为2950万美元）。仅成立1年多就拿到多家投资机构的基金，这从侧面说明公司的发展前景和技术实力；

② 提供的产品较为丰富，官网展示了1年的时间就发布了5种合规性产品。从官网介绍，公司目前拥有130名员工，同时在人员扩充中，开发迭代速度非常快；

③ 产品可快速组合为解决方案，目前提供CCPA、GDPR和LGPD（巴西隐私法规）的合规性方案；

④ 公司的技术创新能力不可小觑。公司掌握了People Data Graph自主技术，将多个分散的个人信息关联到同一个数据主体中，在学术称为“实体识别”问题，这是一个棘手的技术难题，Securiti.ai宣称可将云、数据库、大数据系统等异构数据源关联识别出来，这是一大创新；另外公司利用NLP技术，利用聊天机器人Auti提供友好、不枯燥的处理辅助功能。

Securiti.ai凭借实用的、自动化的合规遵循解决方案、以及不可小觑的创新实力，同时在CCPA今年实施的热点背景下，笔者看好Securiti.ai，让我们拭目以待！

参考链接

[1] https://www.crunchbase.com/organization/securiti-ai#section-overview.

[2] SECURITI.ai Selected as Finalist for RSA Conference 2020 Innovation Sandbox Contest.https://privaci.ai/press-release/securiti-ai-selected-as-finalist-for-rsa-conference-2020-innovation-sandbox-contest/.

[3] Securiti.ai Homepage. https://securiti.ai/.

[4] 2019 网络安全观察. http://blog.nsfocus.net/wp-content/uploads/2020/01/2019-Cybersecurity-Insights.pdf

[5] https://iapp.org/resources/article/ccpa-readiness-survey/

[6] PrivacyOps book. https://www.privaci.ai/request-book.

NSFOCUS

解决方案

绿盟数据安全解决方案

关键词：数据安全、数据梳理、数据生命周期、个人信息保护、追踪溯源、安全运营

摘要：绿盟科技从数据安全建设顶层设计出发，提出“一个中心，四个领域，五个阶段”的数据安全体系建设思路。以数据安全防护为中心，在组织建设、制度流程，技术工具和人员能力4个领域同时开展建设工作，通过“知、识、控、察、行”五个步骤进行数据安全落地建设。

序

《中华人民共和国网络安全法》从发布至今对网络数据和个人信息的保护得到了提升，并促进了经济社会信息化健康发展。

随着党的十九大会议的召开，建设网络强国、数字中国、智慧社会成为全国各地各行各业的重要工作，数字化转型势在必行。随着新型基础设施的推进，数字化世界已经逐渐形成。数字化让办公、学习、生活变得简单、快捷与方便，与此同时数据在集中与共享中不断被放大，数据安全风险在应用中不断滋生。网络和信息安全保护已经成为必然，其中数据安全是重中之重。

近几年，我国就数据安全依法出台多项新政策，2020年《中华人民共和国数据安全法》（草案）和《中华人民共和国个人信息保护法》（草案）相继在人大网上对公众征求意见，草案中为数据安全工作指明了发展方向，提供了法律依据，对整个信息安全产业都带来了积极的影响，让数据安全有法可依、有章可循，为数字化经济的安全健康发展提供了有力支撑。另外对外征求意见的《数据安全管理办法》《个人信息出境安全评估办法》，已发布的《个人金融信息保护技术规范》《信息安全技术个人信息安全规范》《儿童个人信息网络保护规定》《网络安全等级保护制度》等标准与规范都在各自领域发挥着安全作用。

1 数据安全建设工作难点

随着云计算、大数据、物联网、移动互联网、人工智能等新技术的发展，网络边界被不断打破，数字双生、敏捷创新、安全合规驱动快速转型，社会和企业都在面临数字化的转型带来的数据安全风险。

近年来数据泄露的安全事件频发，国家和机构对数据安全的重视程度不断提高，数据安全已经与关键信息基础设施一并成为影响国家稳定、民生安全及社会安全的关键因素。

1.1 数据安全体系建设不完善

- **传统信息安全体系无法保护数据安全：**有别与传统信息安全防护体系，由于数据安全防护体系将保护对象聚焦在“数据资产”这样的无形资产上，数据资产的机密性、完整性以及可用性与硬件资产存在着巨大差别，这导致传统信息安全防护体系通常不具备对数据安全的有效保护能力。
- **静态防护策略无法保护数据安全：**通常一个信息系统中的硬件资产数量是有限的，且在没有重大的系统变更时不会发生显著变化，所以传统信息安全体系的安全策略的设计思路往往是静态的。而随着大数据技术的广泛应用，以及移动互联网应用的蓬勃发展，企业数据存储、处理平台所承载的数据量正在以极快的速度爆炸式增长，若仍以静态的视角看待数据资产势必无法应对数据量急剧增长带来的数据泄漏、数据损坏、数据篡改以及对数据主体造成影响等安全问题。并且由于数据资产对流动性的要求，仅考虑当前主体的静态防护策略显然无法有效保证数据的安全。
- **数据资产的权责不一致：**数据通常来自于企业的业务部门，在业务部门使用，并且数据的所有权也常常属于业务部门，但由于数据安全策略有时会限制业务部门对数据使用的权力，而数据安全体系建设工作由安全部门主导，数据安全防护体系的建设会很有可能受到来自于业务部门的阻力，数据安全体系建设工作推动困难。

1.2 数据安全体系建设目标模糊、建设步骤不清晰

- **缺少数据安全体系建设指导方针：**数据资产在许多环境下对可用性的要求极高，并且由于数据资产对流动性的依赖，如何在保障数据可用性与流动性的前提下落实对数据机密性与完整性的保护是企业所面临的重要问题。
- **缺乏数据安全体系建设经验：**由于数据安全体系建设与传统信息系统安全建设存在着保护范围不同、保护对象不同、安全策略类型不同以及安全建设思路不同等差异，对于企业来说数据安全建设是一项全新的课题。
- **缺少合适的建设指导：**由于我国的数据安全研究正处在逐步推进的阶段，暂时缺少直接有效的指导标准或行业最佳实践帮助企业明确数据安全体系建设的方法与步骤。
- **缺乏有经验的数据安全人员：**由于我国数据安全建设工作正处在起步

阶段，企业安全团队缺少有数据安全经验的人员，这导致数据安全建设难以有效的执行。

2 数据安全体系建设思路

绿盟科技从数据安全建设顶层设计出发，提出“一个中心，四个领域，五个阶段”的数据安全体系建设思路。以数据安全防护为中心，在组织建设、制度流程，技术工具和人员能力4个领域同时开展建设工作，通过“知、识、控、察、行”五个步骤进行数据安全落地建设。

在数据安全体系建设中，组织建设、制度流程，技术工具，人员能力，四个领域都需要同步开展建设工作，组织层面，决策层、管理层、执行层必须在数据安全建设领域达成一致，数据安全建设工作必须得到组织高层的支持。组织高层在数据安全领域的战略目标应该能够被管理层和执行层实现。管理是技术的运营依据、技术是管理的落地保障。所以两者要相辅相成，缺一不可。

绿盟科技借鉴了Gartner的数据安全治理框架，“知、识、控、察、行”的绿盟数据安全治理方法论应运而生，让数据安全落地有声。

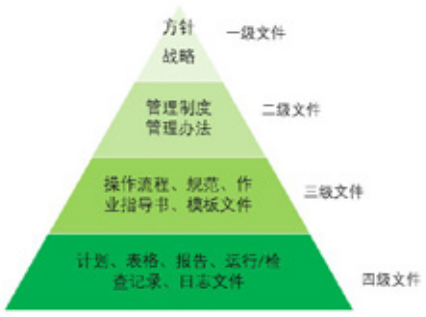
- 知：分析政策法规、梳理业务及人员对数据的使用规范，定义敏感数据；
- 识：根据定义好的敏感数据，利用工具对全网进行敏感数据扫描发现，对发现的数据进行数据定位、数据分类、数据分级。
- 控：根据敏感数据的级别，设定数据在全生命周期中的可用范围，利用规范和工具对数据进行细粒度的权限管控。

- 察：对数据进行监督检察，保障数据在可控范围内正常使用的同时，也对非法的数据行为进行了记录，为事后取证留下了清晰准确的日志信息。
- 行：对不断变化的数据做持续性的跟踪，提供策略优化与持续运营的服务。

3 数据安全体系建设流程

3.1 组织建设与业务数据梳理

在组织与制度设计方面，业务部门要深入参与数据资产梳理以及分级分类工作之中，因为只有业务部门是最了解数据价值与重要性的。因此需要自上而下形成高层牵头，横跨业务部门与安全部门的组织架构。由信息安全管理团队和数据业务管理团队共同商讨建立数据安全制度流程体系。制定好制度体系应该以文档化的方式进行落地管理并严格执行，这样才能更好的开展后续建设工作。



基于业务特点进行数据分类、数据分级。数据分类分级的准确清晰，是后续数据保护的基础。由于数据类型不同，对企业影响不同，我们建议根据《中华人民共和国网络安全法》《个人金融信息保护技术规范》的要求对个人信息和重要数据分开进行评估与定级，再按照就高不就低的原则对数据条目进行整体定级。

3.2 数据全生存周期安全风险评估

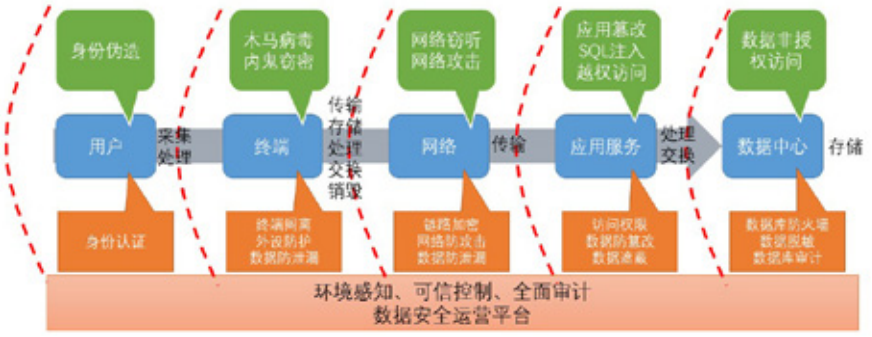
结合数据分类分级要求，辨别人、环境、数据的风险。敏感数据发现与数据风险评估的工作要结合人工服务和专业工具共同完成。

绿盟科技通过对《信息安全技术 数据安全能力成熟度模型》和《个人金融

信息保护技术规范》的研究，整合出一套符合金融行业的数据全生命周期安全风险评估方法，数据生命周期安全风险评估应从通用安全和各阶段安全两个层面进行数据风险检查，了解信息系统总体安全风险状况，对脆弱性的所有方面统一进行分析和评估，并提出整改意见，帮助客户建立快速响应机制，及时有效完成数据安全风险修复工作。

3.3 数据安全纵深防护

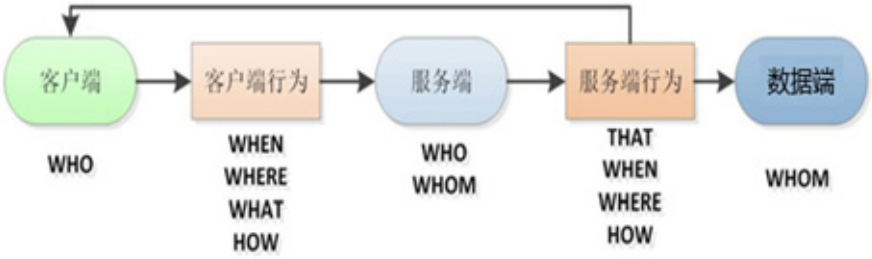
针对数据安全的风险，应以数据为中心，对业务、网络、设备、用户采取“零信任”的态度，既然每个环节都不可信，那么管控手段就要覆盖全部环节，任意环节失信后都能实现熔断保护。



用户侧、终端侧、网络侧、业务侧，以及数据中心，都要做好安全防护措施，外防攻击防入侵防篡改，内防滥用防伪造防泄露。最关键的是，对全部纵深防护环节进行整体控制，实现环境感知，可信控制和全面审计。整合多层次的纵深防护，及时发现问题，及时阻止安全问题。

3.4 敏感数据监察分析

敏感数据监察分析、发现安全问题与异常事件。从用户、资产和数据的行为模式出发，利用5W1H分析模型来进行敏感数据行为分析，基于行为模式发现数据异常事件。

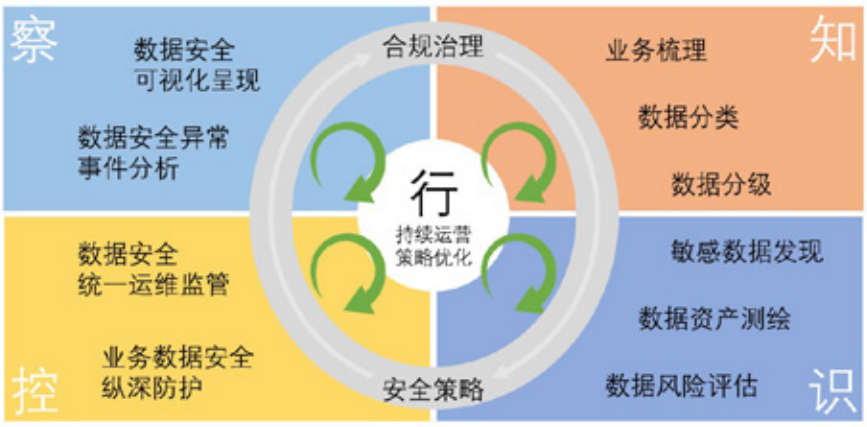


基于历史的可信访问行为提取访问规则，利用各类算法进行行为聚类，形成可划分的访问行为簇并可视化呈现。通过这种图谱分析与可视化展示让管理者对于敏感数据访问情况，由一无所知转变为可视可管。

3.5 优化改进与持续运营

随着业务的运行，数据也在不断变化，因此安全也要不断优化。为了应对变化，必须对数据安全策略进行持续的优化改进与监督运营。

合规要求指导安全策略的设置，安全策略支撑合规治理要求的落地，二者相辅相成，配合上持续优化改进运营的“知识控察行”体系，实现持续自适应的数据安全防护能力。



4 总结

绿盟数据安全解决方案提供了自上而下的数据管理体系，从数据治理到合规监管，从及时预警到风险态势，全方位的解决个人信息和重要数据在企业内的数据安全问题。

随着资产数据化和数据资产化，数据安全成为从大数据时代到云时代发展转变的产物，数据安全的最终落脚点在于数据应用和价值实现，从数据管控向数据价值转变，实现数据驱动业务发展。目前国内数据安全仅在法律法规层面有了方向性的指引，尚缺乏可执行的监管标准和业界最佳实践。必须不断从理论和实践层面完善数据安全建设水平，打造成熟的数据安全建设体系最佳实践，让数据发挥最大价值，从而推动市场经济高速发展。

金融行业数据治理方案

亿赛通 安全服务总监 李迪

关键字：数据安全、数据治理、数据资产、分级分类

摘要：亿赛通根据十六年数据安全实践经验以及对金融行业的深刻理解，提出金融行业数据治理方案，以数据安全防护为核心，围绕数据生命周期，从组织建设、制度流程、技术工具和人员能力等四个方面进行建设，在实施层面按照需求梳理、分级分类、策略制定、技术落地、优化改进五个方面进行数据治理工作的落地开展。

一、 金融行业数据安全背景分析

金融行业作为国家的经济重要领域，数据资产庞大，使用角色繁杂，数据共享和分析的需求强烈，但目前金融机构在数据管理方面仍存在较多问题，数据处理过程中大量的用户信息及用户业务使用信息等个人隐私数据管控机制不足，面临违规越权使用或被用于非法用途等数据泄漏安全风险，对员工有意或无意的敏感数据泄漏缺乏检测与防护手段。

近几年主管部门相继出台了多项规章制度，《网络安全法》于2017年6月1日起施行，银保监会于18年5月发布《银行业金融机构数据治理指引》，证监会于2018年9月发布《证券期货业数据分类分级指引》，从监管层面不断完善数据安全工作，指导金融机构加强数据治理，提高数据质量，以数据驱动金融机构发展。

二、金融行业数据治理方法论

亿赛通数据治理专业服务以数据安全防护为核心，以合规与业务需求为导向，围绕数据生命周期，从组织建设、制度流程、技术工具和人员能力等四个方面进行能力建设。

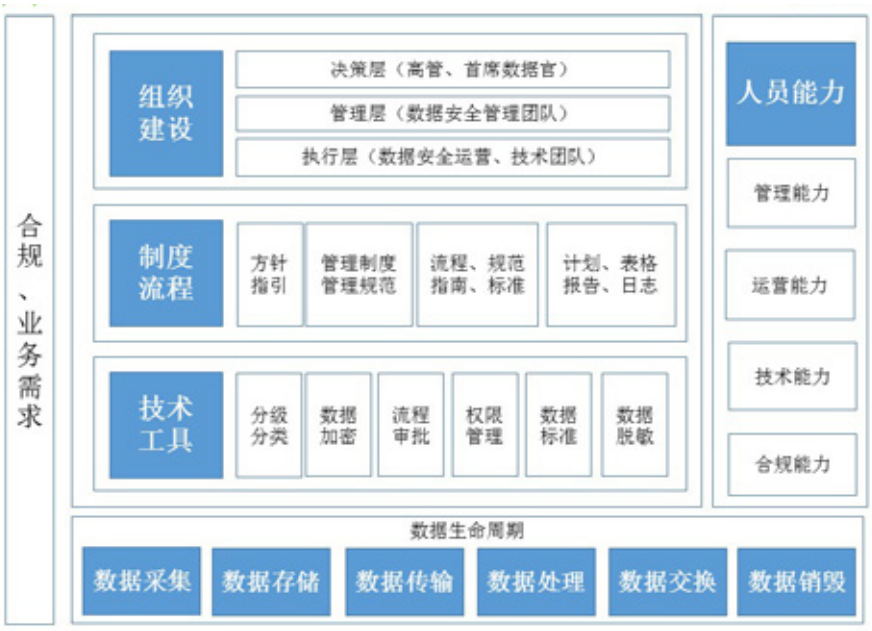


图2.1 数据治理体系架构

2.1 合规和业务需求

在合规监管制度层面，银证监会已经下发了《银行业金融机构数据治理指引》，从业务需求方面，数据向第三方平台共享的安全管控需求，以及数据大集中资产梳理的业务需求驱使推动数据治理建设开展。

2.2 组织架构

传统金融行业安全均由科技部门负责，随着数据治理工作的深入开展，业务部门要深入参与数据资产梳理以及分级分类工作，因此原有的组织架构和项目模式无法支撑数据治理的深入开展，需要自上而下形成高层牵头、跨业务部门、数据全覆盖的组织架构。

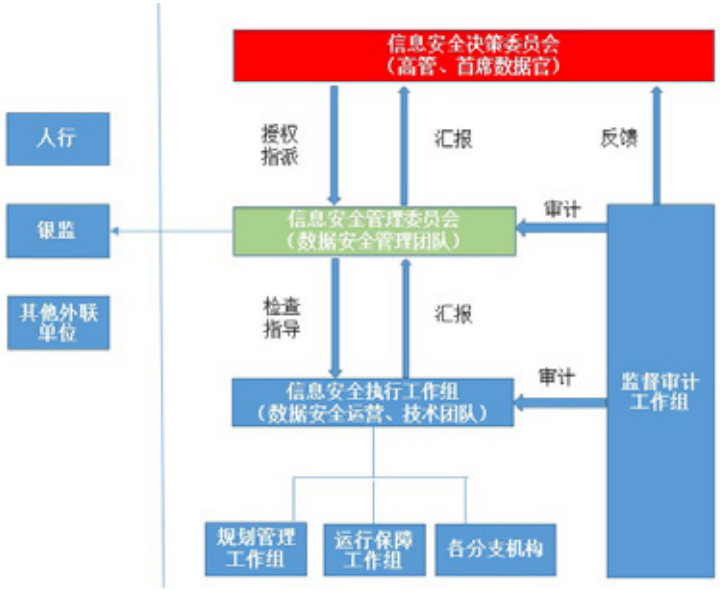


图2.2 数据治理组织架构

2.3 制度流程

目前金融行业大多有较完整的安全规范，如分级分类规定，保密规定等，但一方面没有独立的数据安全规范，可执行性不强，另一方面缺乏技术监管手段，落地执行较难。因此需要制定独立的数据安全管理文件，按不同级别分期建设，逐步落地。

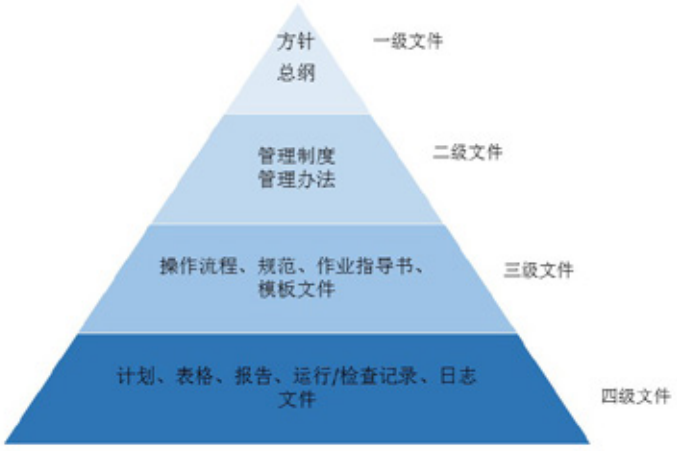


图2.3 数据安全管理制度体系

2.4 技术工具

数据安全项目真正落地执行不仅需要管理制度的规范，更需要技术工具的管控。根据数据分级分类进行安全环境保障、边界管控及合规监管，保障数据的保密性、完整性和可用性。

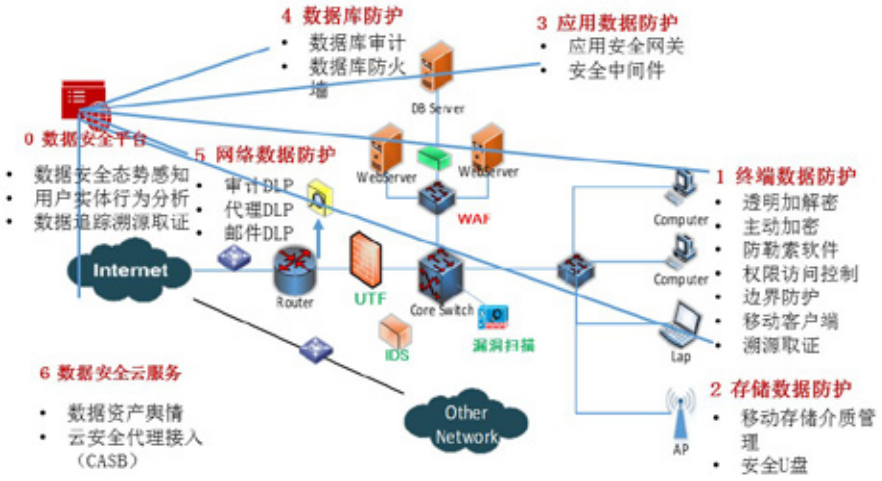


图2.4 数据安全技术工具

2.5 人员能力

传统安全人员的技术能力大多以网络安全和信息安全为基础，而在数据安全层面需要既懂金融业务，又懂数据安全体系的复合型人才，对数据治理人员的培养和管理制度的宣贯需形成常态化机制，提高数据安全人员能力。

三、金融行业数据安全治理实施路径

亿赛通数据安全治理服务按照业务需求、分级分类、策略制定、技术落地、优化改进五个步骤进行数据安全实施工作。

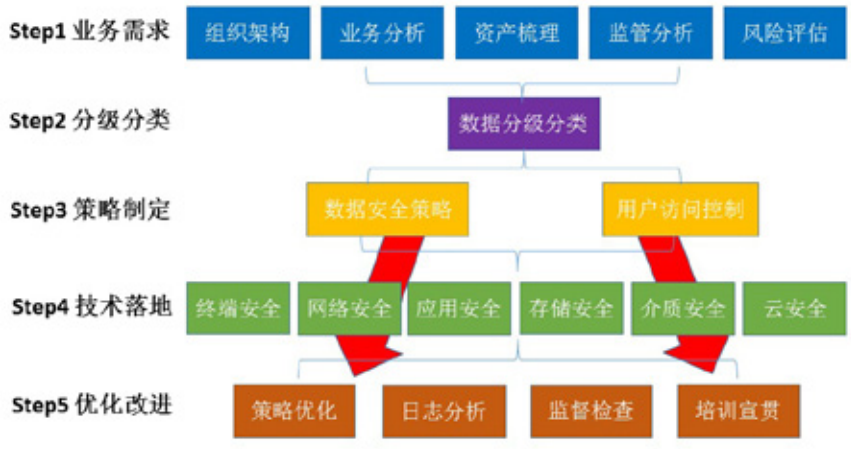


图3.1 数据治理实施路径

3.1 确定业务需求

开展数据治理首先要在确定组织架构的基础上，通过资产扫描工具结合人工业务调研，对企业数据资产进行全覆盖梳理。同时进行全生命周期风险评估和监管政策对标分析，以此确定业务需求和目标。

3.2 数据分级分类

分级分类是数据治理的前提，也是工作量最繁重的环节，在数据资产全覆盖梳理的基础上，首先对条线进行业务细分，确定管理主体和数据治理覆盖范围。其次根据业务调研结果，进行数据资产的数据归类，确定数据类别，完成数据分类工作，然后按照数据损坏丢失可能造成的影响程度进行数据定级，基于数据分级分类对不同级别的数据实行差异化安全控制手段。

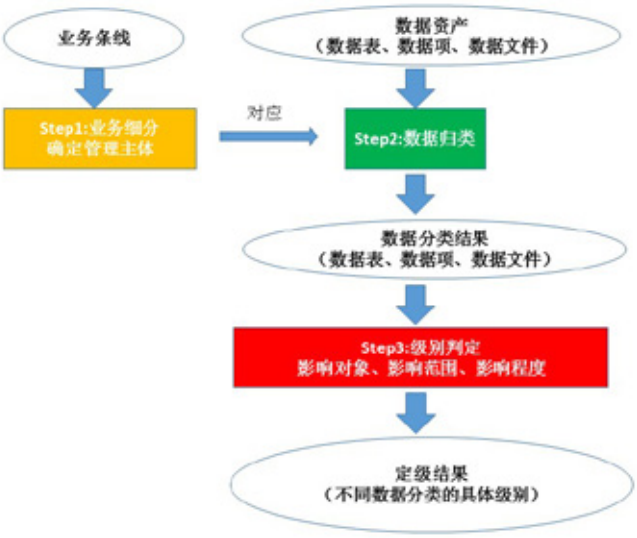


图3.2 分级分类流程

3.3 策略制定

在数据治理的策略制定上，从用户权限控制和数据安全策略两个方向考虑如何实施数据安全治理，在针对“人”的权限控制上，要明确数据的访问者、访问对象、访问行为，尤其在对第三方开放数据共享时，要严格控制系统开放的权限;在针对“数据”的安全防护上，要基于不同级别的数据制定有针对性的数据安全策略，对核心商密和普通商秘数据进行加密管控，对内部公开数据进行安全审计管控，形成整体化全生命周期的数据策略体系。

3.4 技术落地

亿赛通针对数据安全治理提供全生命周期安全管控平台，可以对客户的主机和服务器数据进行全面扫描和梳理，智能识别用户数据安全资产。根据分级分类的落地应用进行标密定密，对于核心数据资产进行加密防护。通过在终端、网络、应用、介质等数据传输通道的全面监控，建立数据安全边界，将技术工具与管理体系有效结合，实现对用户核心信息资产的全方位保护。

图3.3 全生命周期数据安全技术防控手段

在数据安全防控的基础上，可通过数据安全稽核工具进行事后审计，对企业内核心数据资产进行分布统计、合规检查，并对违规行为进行趋势分析和合规预警。打造“事前防御、事中控制、事后审计”的全方位防护体系。

3.5 优化改进

数据安全建设需要长期不断进行管理和技术的调整优化。数据安全平台管控策略要随管理制度细化不断完善优化，也要进行各部门的自查和监督部门的结果性审计检查，同时要做好公司内的培训宣贯，规避常见办公安全风险，贯彻数据安全管理规定。

四、安全治理案例

4.1 Q银行

亿赛通协助Q银行开展业务梳理、资产梳理和分级分类工作，在管理方面制定了数据资产安全管理规章制度，在技术方面建设数据安全管理平台，对办公网内的非结构化电子文档数据共享和移动存储进行加密防护、权限管理和外发管控。

图4.1 Q银行数据安全管理平台

在行内开展安全运营，对内网终端进行全方位审计，并运用大数据分析技术形成信息安全态势分析，提供信息安全事件追踪溯源，统计分析，管理控制等辅助决策技术手段。

Q银行科技处与亿赛通共同完成的《城商行数据安全管理平台的研究及实践》课题获得2018年度银行业信息科技风险管理课题研究成果二类奖项。

4.2 Z银行

亿赛通协助Z银行对总行及全国各分支行进行数据治理工作，进行关键数据梳理、风险评估、制定分级分类标准及相关管理规定，从管理和技术上实现对重要数据资产的分级防护，打造全行数据安全防护体系。

图4.2 Z银行数据治理方法

在项目中利用态势感知技术实现可视化管理、风险预警和溯源取证。实现“授权用、带不走、无法读、留痕迹”的总体目标。

五、结语

数据治理是随着金融行业资产数据化和数据资产化，从大数据时代到云时代发展转变的产物，企业数据治理的最终落脚点在于数据应用和价值实现，从数据管控向数据价值转变，实现数据驱动业务发展。目前国内的数据治理方兴未艾，仅在法律法规层面有了方向性的指引，尚缺乏可执行的监管标准和业界最佳实践。亿赛通将不断从理论和实践层面完善数据治理水平，打造成熟的数据治理业界最佳实践，为金融客户数据安全保驾护航。

绿盟科技数据安全咨询服务介绍

安全服务部 贾晓萍

关键词：数据安全、个人信息安全、咨询服务

摘　要：绿盟科技根据国内外数据安全政策法规及相关标准要求，并借助已有的成熟技术评估服务流程体系及评估工具，为客户提供数据安全专项评估、数据安全治理、数据安全认证及数据安全整体防护方案等适用于多种数据安全场景的咨询服务。

随着社会的进步和科技的发展，信息已经成为我国实现经济转型升级的基础性资源，数据安全则是信息化持续推进的基本前提，但当前的数据保护情况不容乐观，数据泄露、数据滥用、个人信息交易等现象时有发生，数据安全问题日渐凸显，数据安全保护已经成为影响国家安全、社会秩序以及公民利益的焦点问题。

根据国内外数据安全政策法规及相关标准要求，并借助绿盟科技成熟的技术评估服务流程体系及评估工具，通过数据安全专项评估、数据安全治理、数据安全认证咨询及数据安全整体防护方案等多方面咨询服务，为客户提供适用于数据安全多种场景的咨询服务。

一、数据安全咨询服务介绍

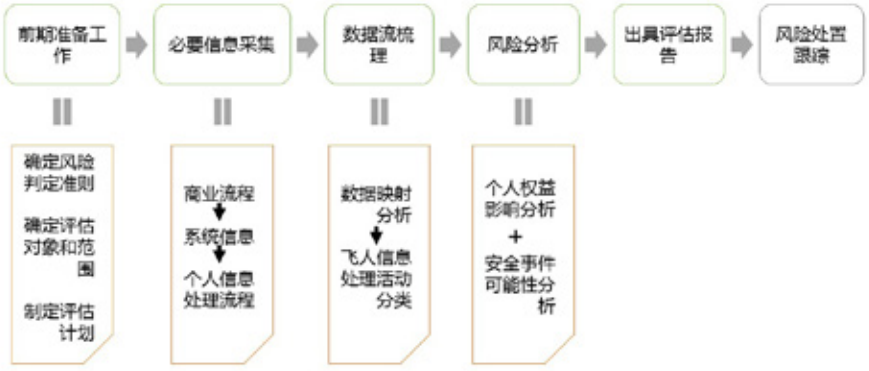
2.1 数据安全专项评估服务

(1) 数据梳理咨询

绿盟科技数据梳理咨询包括数据分类分级和数据映射两部分服务，一方面，通过对数据的识别与收集，对数据进行分类与分级，为数据的分类分级管控做铺垫，另一方面，对收集到的数据进行数据映射，协助客户清晰了解数据的分布状态，为后期数据管控合规性评估做铺垫。

(2) 个人数据影响评估服务

个人信息安全影响评估服务参考《信息安全技术 个人信息安全影响评估指南（征求意见稿）》，通过数据映射分析对客户个人信息处理行为进行梳理与分类，并对客户个人信息处理行为中的个人信息安全风险从危害性和可能性两个维度进行评估，来实现对客户个人信息安全影响的评估工作。根据评估结果，绿盟科技还将提供风险处置的建议，并在需要的情况下对风险处置的结果进行跟踪。



(3) 数据泄露安全评估

面对严峻的数据泄露形势，为积极防范客户由于外部及内部原因导致的数据泄露行为发生，结合其监管机构各项数据安全风险防范工作要求，绿盟科技通过提供基础安全评估工作、业务安全测试、移动APP测试及专项API测试工作，为客户提供数据泄露安全评估服务，进行数据防护手段建设。

(4) 数据生命周期安全评估

绿盟科技数据生命周期安全评估服务从数据生命周期通用安全和各阶段安全两个方面对数据生命周期进行安全评估，以发现客户数据生命周期安全管理及管控措施方面存在的安全问题，提出整改意见。

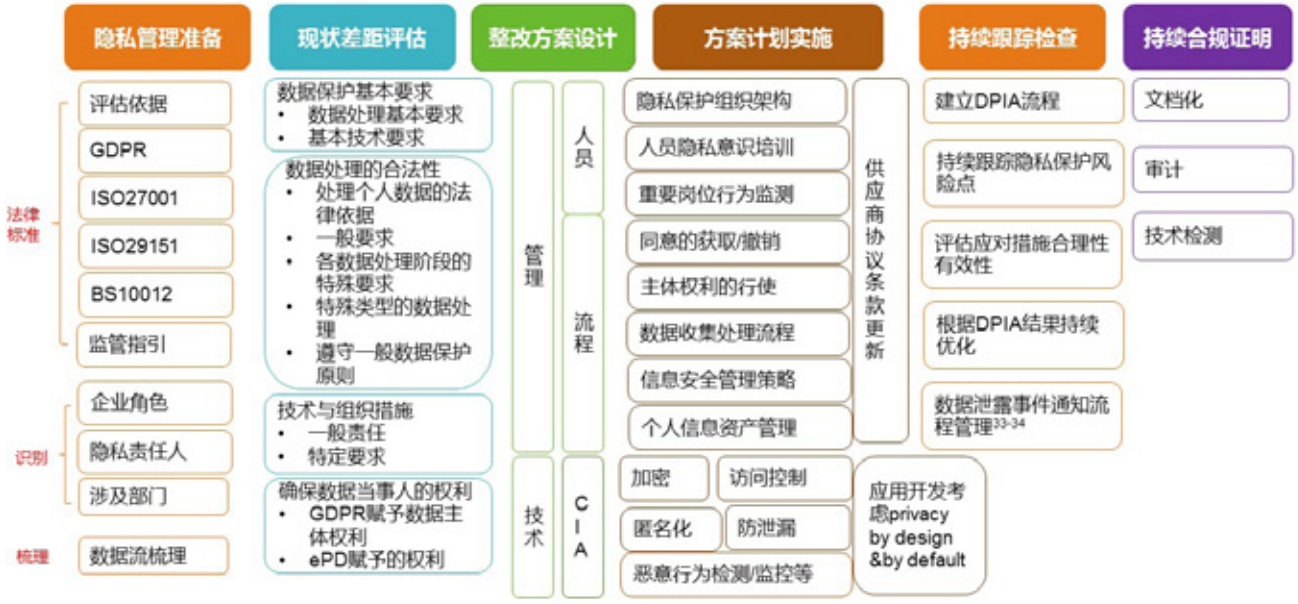
2.2 数据安全治理咨询服务

(1) 通用数据安全治理咨询

绿盟科技通用数据安全治理咨询服务是在对客户数据进行有效理解和分析下，对数据进行不同类别和密级的分类分级工作及数据映射梳理；在对数据分级分类的基础上，了解这些数据在生命周期中的安全管控情况，并辅助以渗透测试、漏洞扫描、配置核查等技术检测工作；最后针对不同数据的安全需求，在满足数据正常使用的目标下，完成相应安全方案的设计、实施及优化服务。

(2) 隐私数据治理（GDPR）安全咨询

隐私数据治理（GDPR）咨询服务，通过详细解读GDPR的相关法规要求，结合国际成熟的信息安全管理体系框架（ISO27001，ISO27018，BS 10012：2017）等，以及先进的技术评估工具，从人员、流程、技术和隐私治理多个方面，帮助客户快速发现不合规的领域并提供应对措施，以督促企业尽快符合GDPR的要求，并通过实施GDPR合规项目，协助客户逐渐形成成熟的隐私数据治理与保护体系。



2.3 数据安全领域认证咨询

近年来，随着对数据安全保护的重视度的提升，数据安全领域的认证也引起了很多企业的关注，绿盟科技目前可提供ISO 29151认证咨询服务，主要参考《ISO/IEC 29151:2017 信息技术-安全技术-个人身份信息保护实践规则》、《ISO/IEC 29100:2011 信息技术 安全技术 隐私框架》要求，对客户相关认证咨询服务，配合客户规范个人信息收集、存储、处理、使用和披露等各个环节中数据操作的相关行为，提高业务流程的安全性和可靠性，降低IT运营过程中的个人可识别身份信息风险，并协助客户获得ISO 29151认证证书。

2.4 数据安全整体防护方案

绿盟科技数据安全整体防护方案从组织建设、人员能力、制度流程、技术工具四方面进行阐述，为客户提供数据安全整体防护方案，协助客户全方位提升数据安全管理与控制水平。

二、数据安全咨询服务客户收益

绿盟科技通过提供数据安全咨询服务，协助客户达到保护数据资产、管理敏感信息、风险规避以及满足政策合规的目的。

大数据安全的解决思路

关键词：大数据平台、数据安全
摘　要：本文从大数据平台下的安全问题，引出大数据安全的法规标准、防护思路及解决方案。

引言

随着互联网、物联网、云计算等技术的快速发展，全球数据量出现爆炸式增长；根据IDC研究的“大数据摩尔定律”表明，人类社会产生的数据一直在以每年50%的速度增长，也就是说，每两年就增加一倍。在大数据不断向各个行业渗透、深刻影响国家的政治、经济、民生和国防的同时，其安全问题也将对个人隐私、社会稳定和国家安全带来巨大的潜在威胁与挑战。

政务信息化的推进，电信、金融、互联网等行业的平台升级，加速推进大数据安全和隐私保护需求。为了应对这些需求，我国正在开展的全国网络安全执行大检查行动中，首次开展针对大数据安全的整治工作，具体包括大数据的采集、传输、存储、处理、交换、销毁等全生命周期的监控与保护。

一、大数据平台下的安全问题

大数据平台涉及到的内容比较广泛，安全问题可以从这5个维度去考虑，安全管理、平台安全、数据安全、运维安全、业务安全。

1、安全管理

安全管理是指大数据平台安全管理方面的要求，包括管理制度、机构和人员管理、系统建设管理、运维管理等内容及配套管理流程。安全防护离不开管理与技术协同，国家、政府、行业自上而下应该有安全管理制度和管理流程，指导具体安全工作的开展和实施。

2、平台安全

平台安全指平台主机、系统、组件自身的安全和身份鉴别、访问控制、接口安全、多租户管理等安全问题，是对大数据平台传输、存储、运算等资源的安全防护要求。企业大多数都使用基于社区化、开源化组件的Hadoop平台，缺乏安全方面的考虑。

3、数据安全

数据属于一种资产，有6个生命周期阶段：采集、传输、存储、处理、交换、销毁；数据安全要保障数据在任何阶段下都是安全的。围绕数据全生命周期考虑数据安全问题，例如：数据采集阶段的分类分级、清洗比对、质量监控；数据传输阶段的安全管理；数据存储阶段的安全存储、访问控制、数据副本、数据归档、数据时效性；数据处理和交换阶段的分布式处理安全、数据加密、数据脱敏、数据溯源；数据交换阶段的数据导入导出、共享、发布、交换监控；数据销毁阶段的介质使用管理、数据销毁、介质销毁等安全问题。

4、运维安全

运维人员的权限相对较大，运维人员直接对数据库进行操作，涉及的数据量非常大，数据的安全难以保障。例如：内部人员的误操作导致数据丢失或不可用，蓄谋恶意行为导致数据泄露。

5、业务安全

业务安全跟业务强相关，跟应用场景和业务流量特征有关，一般的防护手段很难发现，涉及到业务学习和行为分析。例如：缓慢少量攻击、共谋、在噪音中隐身、持续渗漏尝试、长期潜伏者等。

二、大数据安全法规标准

大数据时代是万物互联的时代，数据在共享中体现价值，因此，国内外法律法规也终将完善大数据安全领域的防护和技术要求，助力大数据安全建设。

国家、政府、各行业相继出台大数据平台安全和数据安全相关的国标、行标、企标、地标，推动大数据产业的良性发展。《中华人民共和国网络安全法》、《中华人民共和国计算机信息系统安全保护条例》、《等保2.0》、《个人信息安全规范》、《GDPR》、《电信网与互联网大数据平台安全防护技术要求》等。

三、大数据安全防护思路与解决方案

数据共享是必然需求，大数据安全的防护目标要在保障业务正常的前提下，以合理成本，保护大数据平台下数据的安全。业务需求与风险并存，防护要在业务需求与风险之间寻求平衡，对不同价值和属性的数据，在不同业务需求下，实施不同级别的防护措施，控制防护成本。

1、防护思路

大数据安全防护方案可按层次考虑，平台安全、数据安全、运维安全、业务安全，层层深入，逐步提升安全性。

(1) 平台安全

数据的存储和流转依托大数据平台和各业务系统，平台自身安全是第一

步，通过平台各组件与系统的漏洞扫描管理、规范化的基线核查管理、平台态势感知，确保大数据平台的安全运行。

(2) 数据安全

关注数据的安全存储，数据梳理，掌握数据全景图，让数据风险可量化；关注数据在处理、交换、使用时安全，身份认证、访问控制、数据加密、数据脱敏，防止非法或越权访问数据，对数据访问进行管控、数据审计。

(3) 运维安全

收敛大数据平台的数据访问途径，对运维人员访问大数据平台的操作行为进行操作管控、操作审计。

(4) 业务安全

机器学习建模，对敏感数据的访问行为和敏感业务进行机器学习，对用户行为进行分析，感知和预测业务安全风险。

2、国外厂商方案

Gartner在《Market Trends: Database Security, Worldwide, 2017》报告中列出几个Big Data厂商，我们挑选Informatica和Dataguise这2个厂商的方案进行简单介绍。

(1) Informatica

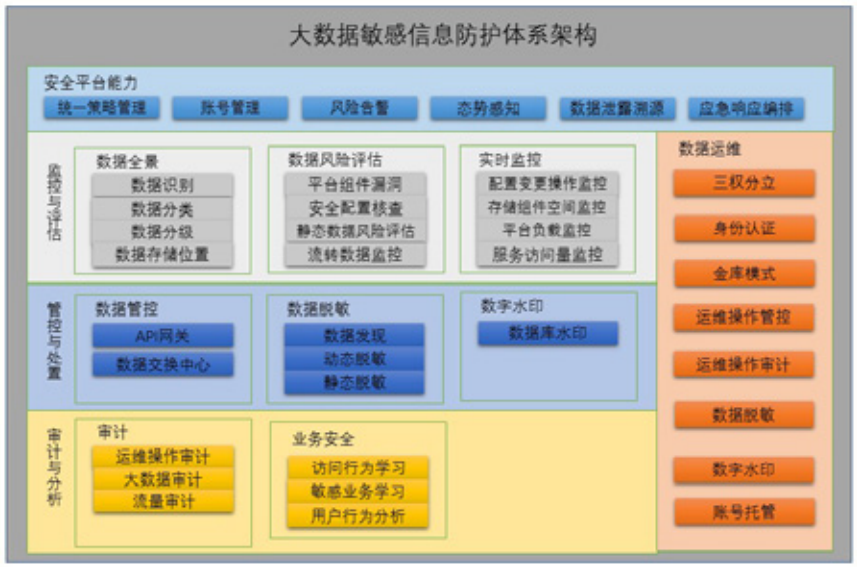
该厂商的大数据安全解决方案支持对结构化、非结构化数据做发现、分类、风险评分，数据访问和操作监控，发现可疑或未授权操作，数据保护，敏感数据扩散跟踪，让风险跟踪处置形成闭环。

(2) Dataguise

该厂商提供全局敏感数据管理解决方案，产品旁路部署在大数据集群边界，可实时检测、审计、保护和监视敏感数据资产。

3、绿盟大数据安全防护方案

绿盟结合数据安全防护思路和实操性，从安全平台能力、运维安全管控、监控与评估、管控与处理、审计与分析这几方面去考虑，设计出一套大数据安全防护方案，覆盖平台安全、数据安全、运维安全、业务安全，提供数据发现、分类、分级、评估、监控、保护、审计、溯源、态势感知一整套大数据安全防护方案。



安全平台对接设备，集中管理、日志收集、智能分析，拥有统一策略管理、账号管理、风险告警、态势感知、数据泄露溯源、应急响应编排能力。

监控与评估，从监控的角度入手，首先对大数据平台上的数据做梳理，数据识别、分类、分级、存储位置定位，生成数据全景图，为数据细粒度的访问授权提供依据，同时对动态数据做跟踪，监控数据流转、使用是否符合预期；从评估的角度出发，结合平台自身的组件漏洞、配置安全性及数据在

平台上的存储、流转情况做综合评估，全面剖析数据风险，量化数据风险，为大数据平台的态势监测与防护提供有力支撑。

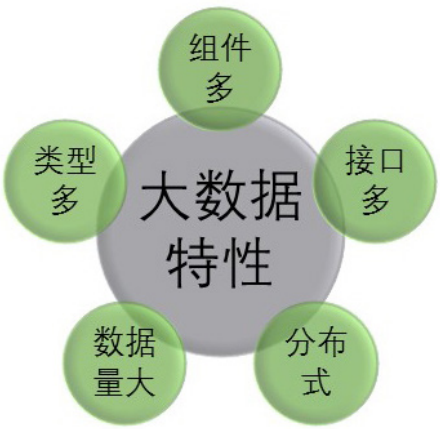
管控与处置，结合客户行业数据特征，提供行业数据分类分级模板；结合客户实际业务需求，灵活提供细粒度访问控制、数据加解密和数据脱敏方案，根据实时应用和业务流量监控，及时处置异常行为，避免进一步风险。

审计与分析，审计指对行为操作和数据访问做审计，为事件问题定位、溯源和大数据分析提供依据；业务安全指基于用户画像和异常行为分析做业务安全风险的感知和预测，及时给出处置策略。

运维安全管控，对平台上所有运维操作进行统一管控和审计，利用运维操作审核机制，防止内部人员私自、独立对平台配置和数据进行操作。（等保2.0要求：大数据平台的管理流量和系统业务流量分离，因此，运维流量和业务流量也要分开做管控）

四、大数据安全面临的挑战

大数据安全与传统数据安全相比，存在一些差异，大数据环境的特点是分布式、组件多、接口多、类型多、数据量大，这些特性给大数据安全引入了技术难点。



主流开源大数据组件二十多款，还有大量第三方封装的组件，不同组件使用的交互接口不同，安全产品面对这么多组件接口，在监控、防护、溯源的方案设计和技术实现上都有难度。

大数据平台要存储和处理的数据量庞大，IDC预计，到2020年全球数据总量将超过40ZB，面对持续膨胀的数据量，安全产品不仅要提高单机产品的处理性能，还要考虑产品扩容和延展性。

大数据平台要存储和处理的数据类型众多，结构化数据、半结构化数据、非结构化数据。要对非结构化数据做识别、分类分级和脱敏处理，有一定技术难度。

五、大数据安全未来发展方向

由于政务大数据覆盖了自然人、法人、企业、政府机构等，同时和医疗、教育、民生服务等各个部门相关；因此，解决了政务大数据安全问题，就能有效解决其他行业大数据安全问题，有力支撑国家治理体系和治理能力现代化目标的实现。从企业层面来看，国家将统一标准规范，避免行业交流繁杂、数据所有权混乱、开发成本高等一系列问题。统一的数据管理平台，统一的数据存储，统一的数据标准，进行统一的数据资产管理，统一进行授权管理，这是未来探索的一个方向。

让安全更有效
绿盟科技安全服务

专业|灵活|高效

咨询服务
安全规划
合规咨询
信息安全管理体系咨询
应急体系建设
……

可管理安全服务
远程安全运维
全评估/测试服务
安全基线服务
应急响应
……

安全评价
外部检查辅导
安全指标体系度量
……

安全研究
渗透测试
源代码审计
业务安全测试
漏洞挖掘
……

教育培训
安全技能培训
安全意识教育
……

THE EXPERT
BEHIND GIANTS
巨人背后的专家

多年以来，绿盟科技致力于安全攻防的研究，为运营商、政府、金融、能源、互联网以及教育、医疗等行业用户，提供具有核心竞争力的安全产品及解决方案，帮助客户实现业务的安全顺畅运行。在这些巨人的背后，他们是备受信赖的专家。

客户支持热线：400-818-6868

NSFOCUS绿盟科技

安全月报

绿盟科技金融事业部出品

主办/绿盟科技金融事业部
地址/北京市海淀区北洼路4号益泰大厦3层
邮编/100089
电话/010-59610688-1159
传真/010-59610689
网站/www.nsfocus.com
客户支持热线/400-818-6868
股票代码/300369
月报电子版下载/http://www.nsfocus.com.cn/research/list_145_145.html