

企业数据安全中的数据脱敏



张弛 阿里巴巴本地生活



目录

CONTENTS

01 相关概念

02 数据脱敏
常见方法

03 企业内部
脱敏场景

04 延伸思考

01 题目

Subject

数据脱敏相关概念



数据脱敏基本概念

Data Desensitization（数据脱敏），

是指在**不影响数据分析结果的准确性**的前提下，对原始数据中的敏感字段进行处理，从而**降低数据敏感度和减少个人隐私风险**的技术措施。

Data Masking（数据脱敏），

是屏蔽敏感数据，对某些敏感信息（比如，身份证号、手机号、卡号、客户姓名、客户地址、邮箱地址、薪资等等）通过脱敏规则进行数据的变形，实现隐私数据的可靠保护。

De-identification（去标识化），

是指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。

Anonymization（匿名化），

是指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关联，且处理后的信息不能被还原的过程。

方法

效果

02 题目

Subject

数据脱敏常见技术方法

常见数据数据脱敏（去标识化）技术

ICS 35.040
L 80



中华人民共和国国家标准

GB/T 37964—2019

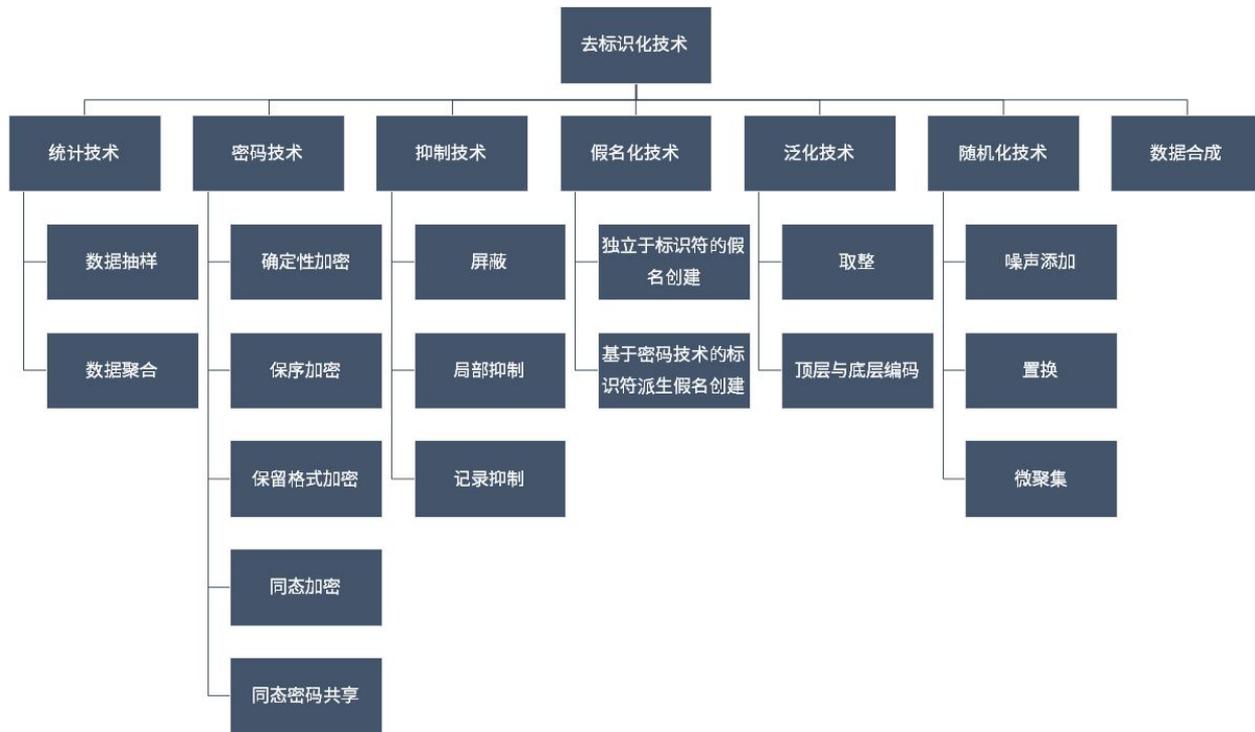
信息安全技术
个人信息去标识化指南

Information security technology—
Guide for de-identifying personal information

2019-08-30 发布

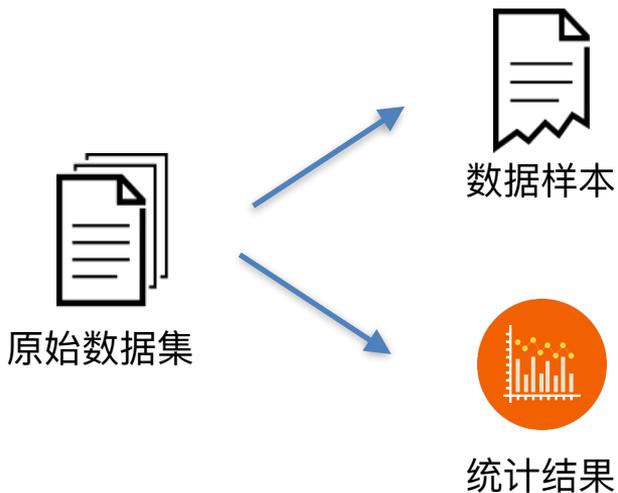
2020-03-01 实施

国家市场监督管理总局 发布
中国国家标准化管理委员会



统计技术

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

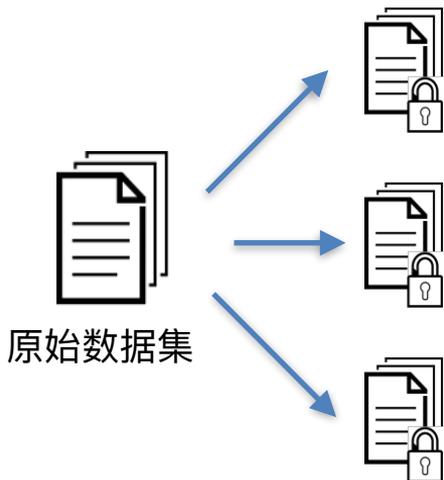


数据抽样是通过选取数据集中有代表性的子集来对原始数据集进行分析和评估的，它是提升去标识化技术有效性的重要方法。

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

密码技术

密码技术是去标识化或提升去标识化技术有效性的常用方法，采用不同类型的加密算法所能达到不同的脱敏效果。



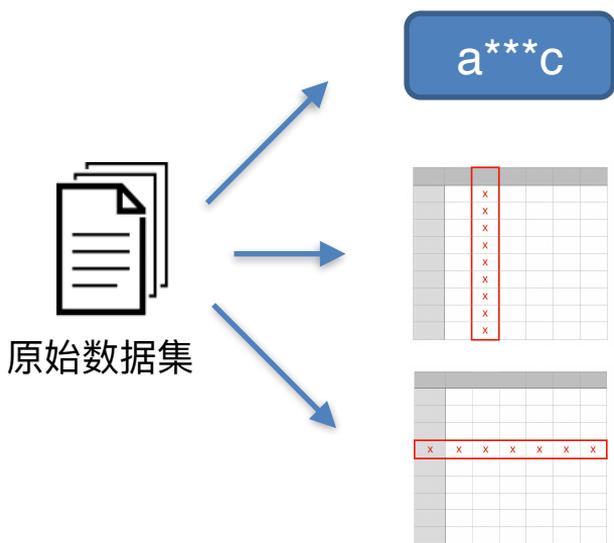
确定性加密，是一种非随机对称加密，常见对id类数据进行处理，可在必要时对密文进行解密还原为原id，但需要对密钥进行妥善保护。

不可逆加密，通常散列（hash）函数对数据进行处理，常见于对id类数据进行处理，不可以直接解密，需保存映射关系，同时因为hash函数特性，会存在数据碰撞的问题，用法简单，不用担心密钥保护。

同态加密，采用密文同态算法，其特点是密文运算的结果解密之后和明文运算相同，因此常见于对数值类字段进行处理，但性能原因，目前未大范围使用。

抑制技术

抑制技术即对不满足隐私保护的数据项删除或屏蔽，不进行发布。



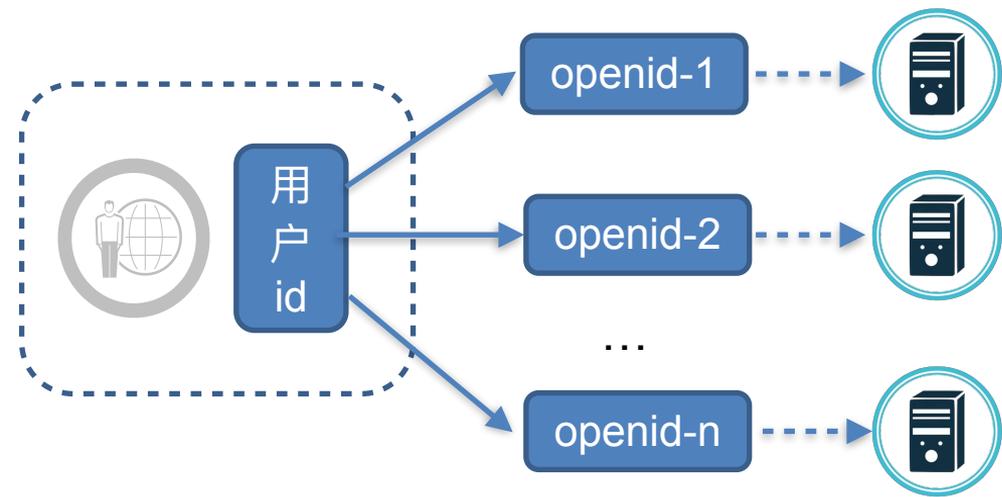
屏蔽，是指对属性值进行屏蔽，最常见的脱敏方式，如对手机号、身份证进行打*号处理，或对于地址采取截断的方式；

局部抑制，是指删除特定的属性值（列）的处理方式，删除非必要的字段；

记录抑制，是指删除特定的记录（行）的处理方式，删除非必要的记录。

假名化技术

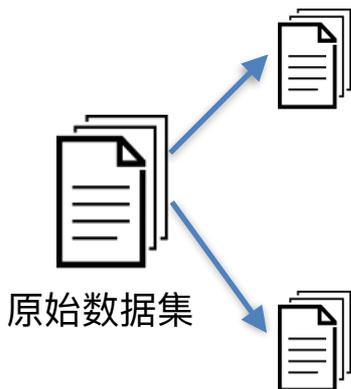
假名化技术是一种使用假名替换直接标识（或其它敏感标识符）的去标识化技术。假名化技术为每一个人信息主体创建唯一的标识符，以取代原来的直接标识或敏感标识符。



- 可以独立生成随机值对原始ID进行对应，并保存映射关系表，同时对映射关系表的访问进行严格控制；
- 同样可以采用加密的方式生产假名，但需为妥善保存解密密钥；
- 该技术广泛使用在数据使用方数量多且相互独立的情况，比如开放平台场景的openid，同样一个用户，不同开发者获取的openid不同。

泛化技术 & 随机化技术

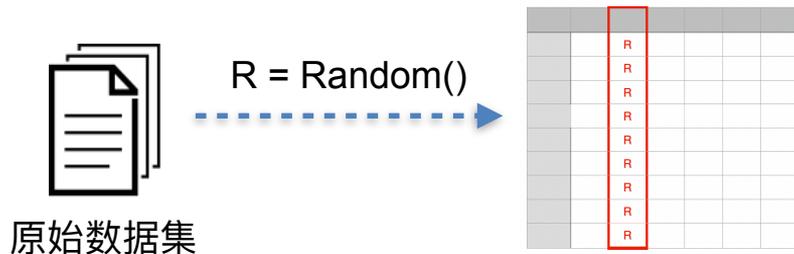
泛化技术是指一种降低数据集中所选属性粒度的去标识化技术，对数据进行更概括、抽象的描述。泛化技术实现简单，能保护记录级数据的真实性，常见于数据产品或数据报告中。



取整涉及到为所选的属性选定一个取整基数，比如向上或向下取证，产出结果100、500、1k、10k

顶层与底层编码技术使用表示顶层（或底层）的阈值替换高于（或低于）该阈值的值，产出结果为“高于X”或“低于X”

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



03 题目

Subject

企业内部常见数据脱敏场景



静态脱敏

适用场景

批量进行脱敏数据

- 用于应用功能测试的测试数据
- 用于模型训练的测试数据
- 数据导出用于终端/离线分析

技术方法

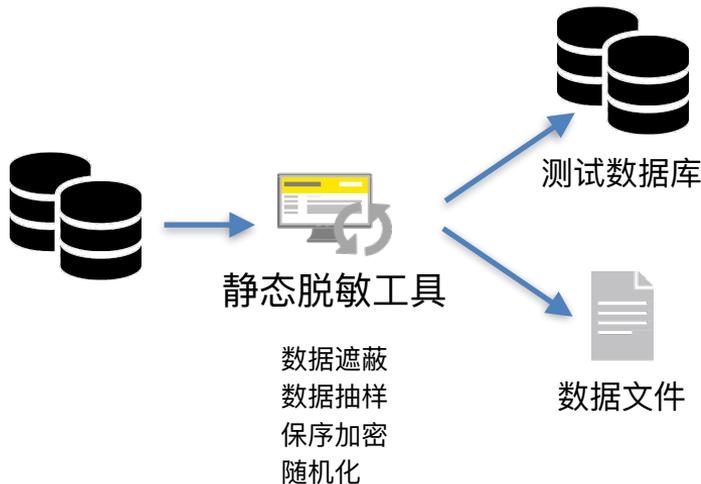
不同用途选取不同的方式或组合

- 应用测试数据：随机化（测试功能），数据遮蔽（对格式无要求）
- 模型训练数据：对于id进行假名化，对于属性进行保序加密，数据抽样
- 数据导出用于终端/离线分析：数据屏蔽、局部抑制（删除列）

落地细节

常见问题和注意事项

- 脚本/工具，基于ETL工具或数据同步工具进行改造
- 字段类型识别，及对应脱敏方式的映射关系
- 通过网络ACL、数据库权限等方式，对数据导出进行收敛



数据库（动态）脱敏

适用场景

技术人员直接操作数据库

- 研发人员的开发调试
- DBA日常数据管理
- 运维人员的基础运维

技术方法

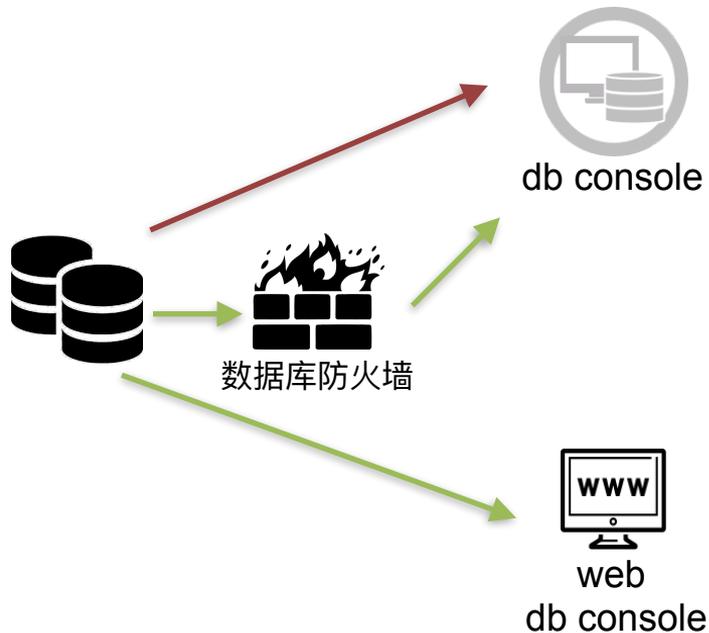
不同用途选取不同的方式或组合

- 最常见的方式采取屏蔽技术，如对敏感字段打*号

落地细节

常见问题和注意事项

- 数据库防火墙：基于数据库协议，对sql语句进行改造，增加脱敏udf；或对返回结果进行转化
- web console：通过应用来访问数据库，在应用上通过抑制技术前端展示脱敏、限制查询条数等
- 必须限制源生db console的使用，可以通过数据库防火墙以及数据库端口的网络ACL。



应用系统(动态)脱敏

适用场景

应用系统的数据脱敏

- 前端页面的敏感数据脱敏
- 数据类接口（API）的透出数据脱敏

技术方法

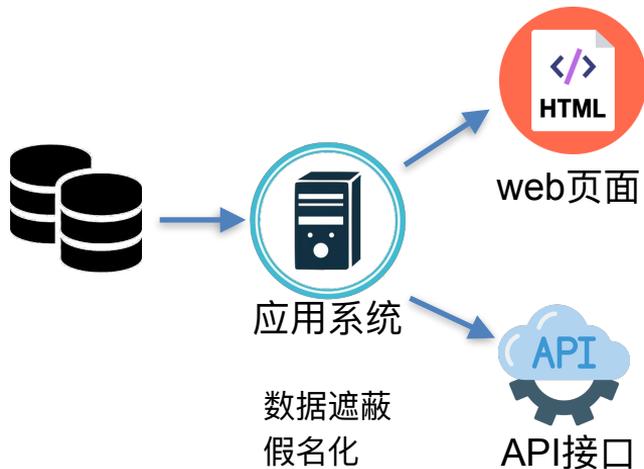
不同用途选取不同的方式或组合

- 对于属性类，可采取屏蔽技术，如对身份证、手机号等在展示时打*号
- 针对ID类的，可采取假名化进行处理

落地细节

常见问题和注意事项

- 脱敏的字段和规则需提前定义，对于前端页面可以在模版中引入脱敏函数；对于API可以通过API网关统一处理；
- 一般比较少用数据库防火墙，会影响应用中的基于真实数据业务逻辑处理，同时性能会是瓶颈；
- 脱敏一定要在服务端进行，前端通过js



大数据平台综合场景

适用场景

大数据平台的各个环节

- ETL过程中的数据抽取和加工
- 面向分析人员（类数据库动态脱敏）
- 结果数据导出（类静态脱敏）

技术方法

不同用途选取不同的方式或组合

- 针对ID类主键，可采取假名化对ID进行归一化或者转换
- 对于非主键的属性类值，可以采用保序加密或者对称加密的方式
- 面向分析人员的分析工具，采用屏蔽的方式实现动态脱敏

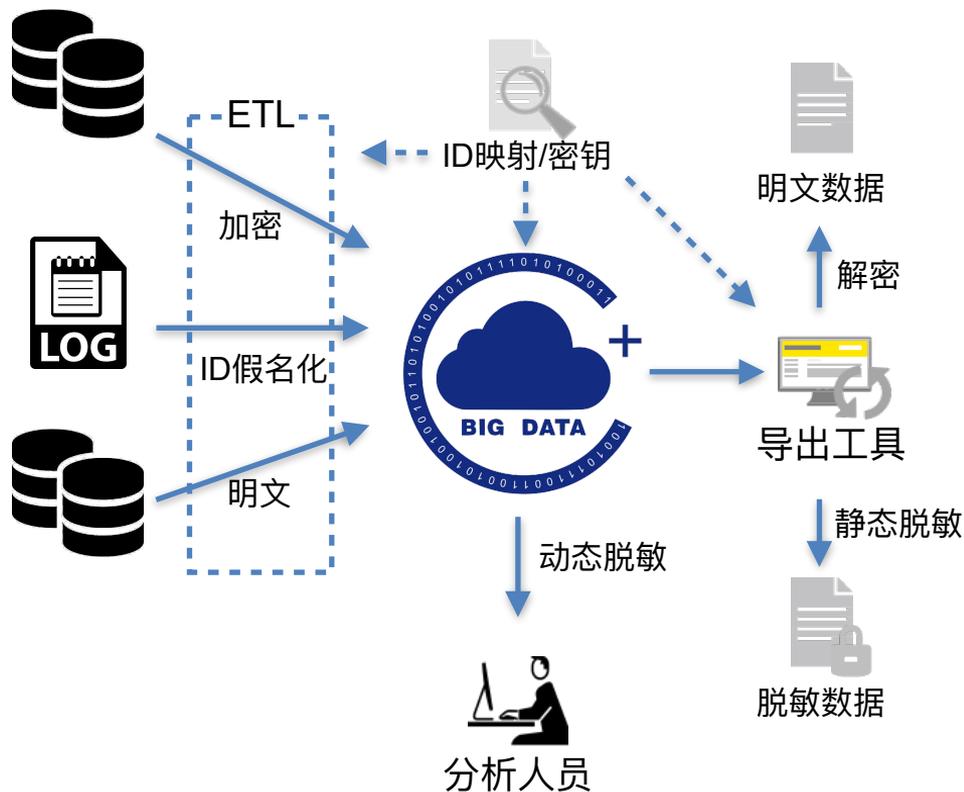
落地细节

常见问题和注意事项

- 需要开展数据分析，比较少用屏不可逆的方法，比如屏蔽，但可以针对分析人员的分析工具进行查询结果数据的进行屏蔽；
- ID映射、加密密钥务必严格控制可访问权限

同样以来对敏感字段类型的识别，并采

脱敏方法。



数据产品&数据报告脱敏

适用场景

数据类应用或编写数据报告

- 内部数据监控类产品或看板
- 对外服务的数据类产品
- 基于数据分析的报告，如业务汇报、项目复盘等

技术方法

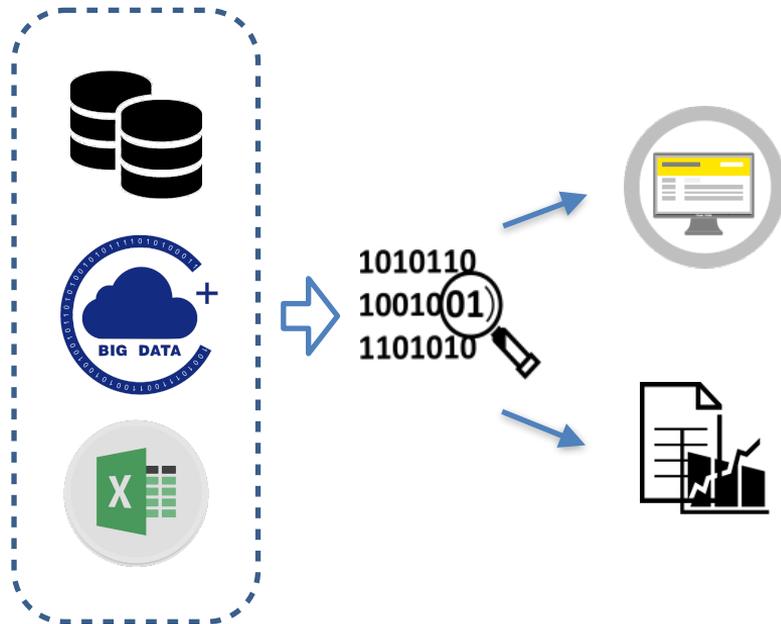
不同用途选取不同的方式或组合

- 常见的采用数据聚合的方式，通过统计分析，通过多种类型的图表呈现排名、变化趋势、占比情况等；
- 对于数值类采用泛化技术，
- 通过数据抽样技术选取特定数据范围

落地细节

常见问题和注意事项

- 尽量避免出现绝对数值，为凸显变化可用同比环比；
- 排名类尽量只保留排名，避免通过占比情况反推大盘数据；
- 数据的抽样范围必须与产品/报告呈现最终呈现的对象相匹配，避免出现反推上一级类目或更大区域的情况。



04 题目

Subject

数据脱敏延伸思考

数据脱敏的延伸思考

产品经理：我需要一把**大铁锤**

安全工程师：大铁锤**太危险**，不可以！

产品经理：那业务跑不通，还要你们安全做什么？

安全工程师：可以考虑给你一把**小锤子或者木头锤子**

产品经理：那行吧，先将就用着

（……一段时间后……）

产品经理：我还是需要一把大铁锤！

安全工程师：……，大铁锤太危险，不可以！

产品经理（内心OS）：

我就只想**在墙上敲个洞**，真费劲！

产品经理：我需要用明文手机号

安全工程师：明文手机号**风险太高**，不可以！

产品经理：那业务跑不通，还要你们安全做什么？

安全工程师：……，**前三后四中间打***，可行

产品经理：那行吧，先将就用着

（……一段时间后……）

产品经理：我还是需要明文手机号！

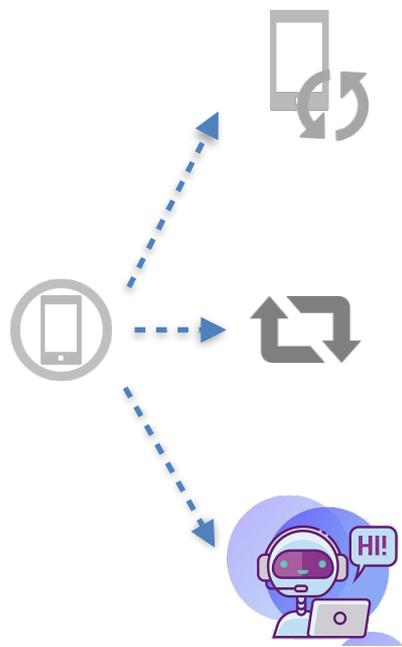
安全工程师：……，明文手机号风险太高，不可以！

产品经理（内心OS）：

我就只想**联系一下用户**，真费劲！

是否还有其他的可能？？？

数据脱敏的延伸思考



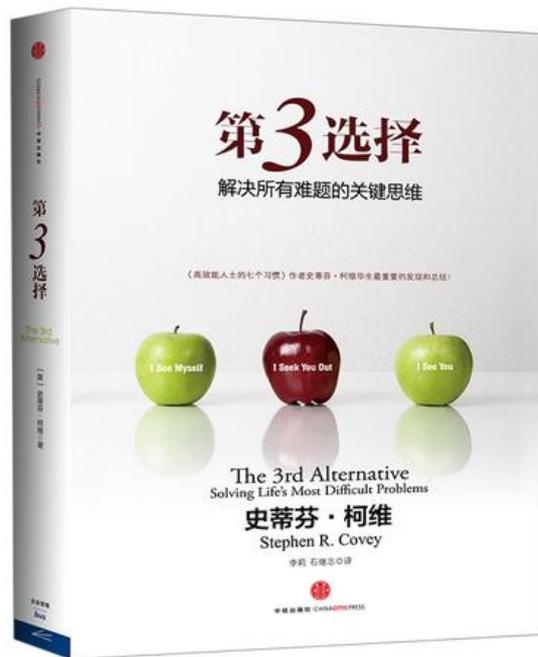
隐私小号，通过另外一个临时性的号码来替换业务过程中的真实手机号，通过对主叫、被叫、信息功能及时效的控制，来减少用户真实手机号的暴露面，进而降低数据泄漏的可能。

语音双呼，通过呼叫服务分别回拨呼叫者和被呼叫者，在不向呼叫者提供真实号码的同时建立双方通话，比如CRM上的呼叫按钮、钉钉智能电话。

智能外呼，通过智能机器人主动发起语音呼叫，并对客户语音识别迅速，能够准确判断出是否为意向客户，适用于数量大、重复、机械的初步筛选，以此减少人的参与，进而降低数据泄漏的可能。

数据脱敏的延伸思考

- 数据给？还是不给？怎么给？这就是数据安全
- 不要在数据本身上过多纠结，请多问一句“why”
- 没有最强大的数据安全方案，只有最合适当下的
- 业务和安全的大部分冲突，归根于彼此不够了解
- 业务和安全之间，尝试一下《第三选择》？



THANKS!

Ending

