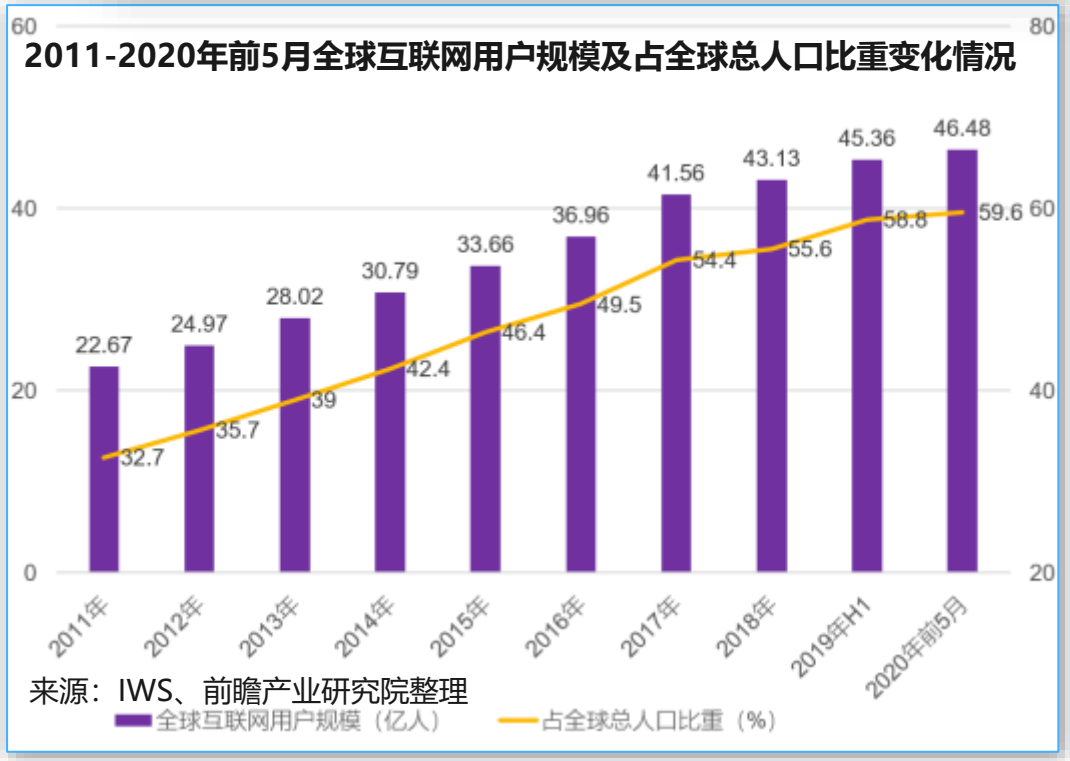
隐私保护

李琦
清华大学网研院

隐私泄露成为大众关注焦点

互联网上的隐私侵犯问题已引发互联网用户的普遍担忧，成为了大众关注的焦点



圆通“内鬼” 泄露40万条个人信息

哔哩哔哩后台源码泄露

以色列640万选民数据遭泄露

简历信息被贩卖

美国金融公司Evite泄露1亿客户的信息

超50万个Zoom账户泄露

中国电信超2亿条用户信息被卖

Facebook用户信息被“剑桥分析”获取

优衣库泄露超过46万名客户数据

国泰航空泄露940万乘客信息

谷歌浏览器大规模用户安全信息泄露

- 截至2020年5月31日，全球互联网用户数量达到46.48亿，占世界人口数量的59.6%
- 2000-2020年，世界互联网用户数量增长了近12倍

- 近年来，互联网隐私泄露事件数不胜数，关于隐私泄露的新闻层出不穷

讨论

畅所欲言

你觉得职业、身份证、手机号码、住址、职业、收入、购物记录、上网记录、朋友圈信息中哪些是你的隐私?

本章的内容组织

第一节
隐私保护技术初探

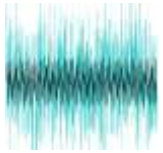
- 网络空间安全中的隐私
- 隐私泄露的危害
- 隐私保护技术介绍

隐私泄露事件造成的危害让人们看到了隐私保护的重要性

第二节
匿名化

- 匿名化隐私保护模型
- 数据匿名化方法

如何安全地发布数据供其他机构研究



第三节
差分隐私

- 差差分隐私基础
- 数值型差分隐私
- 非数值型差分隐私

如何保护统计信息中的个体隐私

第四节
同态加密

- 同态加密基础
- 半同态加密
- 全同态加密

如何安全地将数据委托给数据计算方

第五节
安全多方计算

- 安全多方计算基础
- 百万富翁协议

如何帮助互不信任的参与方进行协同计算

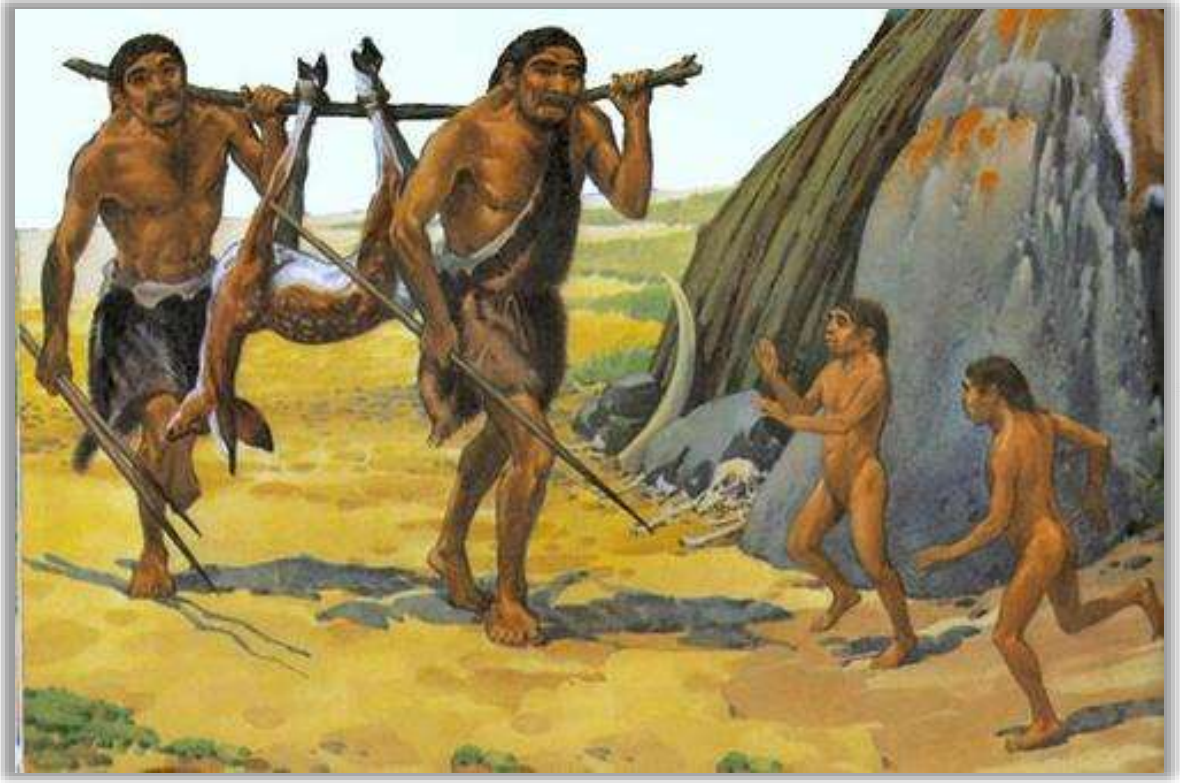
隐私泄露的巨大危害促使人们考虑多种场景下的数据隐私保护问题

第1节 隐私保护技术初探

- 网络空间安全中的隐私
- 隐私泄露的危害
- 隐私保护技术介绍

隐私概念的发展

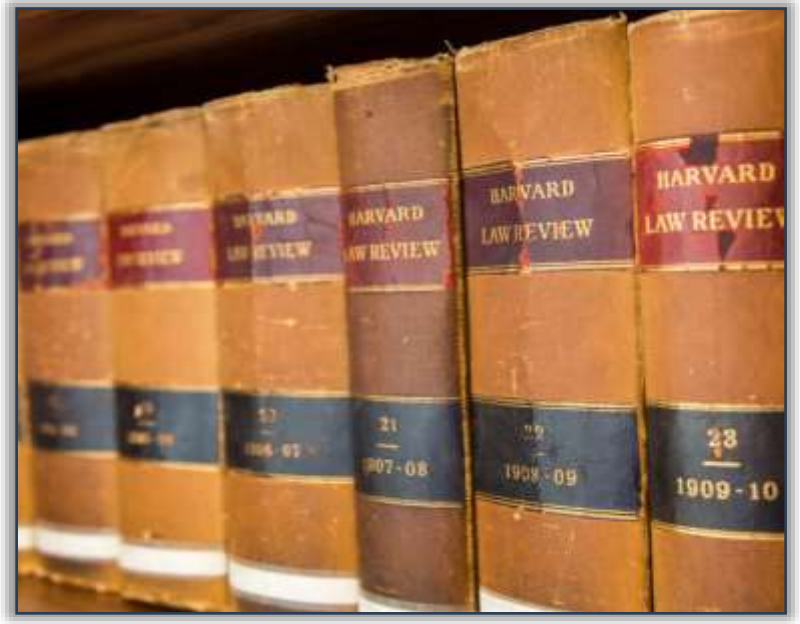
远古时代人类开始用树皮、动物皮毛等来遮挡身体的隐私部位代表了隐私萌芽的诞生，此后，随着社会的发展，隐私以各种各样的形式存在于人们的生活中



隐私权的诞生

- 虽然人类早就意识到了隐私的重要性，然而将隐私作为人的一项权利的意识直到西方自由主义观念盛行后才得到重视
- 在1890年，当时的美国报纸行业为了迎合大众口味，大量报导各种犯罪、丑闻和名人私生活
- 美国学者塞缪尔·沃伦由于不满报纸对其家庭生活的报道，和路易斯·布兰代斯共同执笔在《HARVARD LAW REVIEW》期刊上发表了一篇名为“THE RIGHT TO PRIVACY”的文章，**被公认为是对隐私权的首次提及**

在残酷的市场竞争中，许多报纸为了争夺读者、提高发行量及在竞争中取胜并谋取良好收益，不惜以耸人听闻的手法大量报道犯罪、色情、丑闻和社会猎奇之类题材的刺激性新闻，对之进行夸张渲染。在版面设计上，这些报纸追求视觉效果，编制故作惊人的通栏加黑大标题，标题内容也极尽煽情之能事。采用大幅图片和色彩。以此引起轰动效应，刺激社会神经，吸引读者关注。这类新闻被人们称为“黄色新闻”，这一时期被称为美国新闻事业的“黄色新闻时期”。



THE RIGHT TO PRIVACY

HARVARD

LAW REVIEW.

VOL. IV. DECEMBER 15, 1890. NO. 5.

THE RIGHT TO PRIVACY.

"It could be done only on principles of private justice, moral fitness, and public convenience, which, when applied to a new subject, make common law without a precedent; much more when received and approved by usage."

WILLES, J., in Millar v. Taylor, 4 Burr. 2303, 2312.

THAT the individual shall have full protection in person and in property is a principle as old as the common law; but it has been found necessary from time to time to define anew the exact nature and extent of such protection. Political, social, and economic changes entail the recognition of new rights, and the common law, in its eternal youth, grows to meet the demands of society. Thus, in very early times, the law gave a remedy only for physical interference with life and property, for trespasses *vi et armis*. Then the "right to life" served only to protect the subject from battery in its various forms; liberty meant freedom from actual restraint; and the right to property secured to the individual his lands and his cattle. Later, there came a recognition of man's spiritual nature, of his feelings and his intellect. Gradually the scope of these legal rights broadened; and now the right to life has come to mean the right to enjoy life,—the right to be let alone; the right to liberty secures the exercise of extensive civil privileges; and the term "property" has grown to comprise every form of possession—intangible, as well as tangible.

Thus, with the recognition of the legal value of sensations, the protection against actual bodily injury was extended to prohibit mere attempts to do such injury; that is, the putting another in

生活的强烈和复杂，伴随着逐渐进步的文明，让人们偶尔需要从外界逃离。在不断完善的文化影响之下，人类对公众的感知越来越敏感，所以独处与隐私成为了个体的基本需求；但是，现代企业和发明通过侵蚀个人隐私，强迫个人屈从于心理痛苦与压力，这比让一个人遭受肉体创伤要严重得多

——《THE RIGHT TO PRIVACY》

隐私的定义

- 古汉语：“隐”为隐藏，隐避，“私”为私人、私下
- 汉语词典：不愿告人或不愿公开的个人的事
- 百度百科：与公共利益、群体利益无关，不愿告人或不愿公开的个人的私事

简单来说，隐私就是个人或者团体不愿被他人知晓的信息

网络空间中的隐私

个人数据

网络行为数据

通信内容

保护隐私的法律法规

- 1970年德国:《联邦数据保护法》
- 1974年美国:《隐私权法》
- 1984年英国:《数据保护法》

国外

国内

- 1996年：香港《个人资料（隐私）条例》
- 2017年：《网络安全法》
- 2020年：新版《个人信息安全规范》
- 2021年：《数据安全法》
  《个人信息保护法》

隐私保护领域最受关注的法律：欧盟在2018年生效的GDPR（通用数据保护条例）

- 该法律生效后，即使企业没有在欧洲直接展开业务，也未在欧洲设立任何分支机构，但是只要涉及处理欧洲公民的个人数据，都要遵循该法律条款
- 对违法企业的罚金最高可达2000万欧元（约合1.5亿元人民币）或者其全球营业额的4%，以高者为准
- GDPR生效17个月，欧洲开出约3.7亿欧元罚单

隐私泄露数据统计

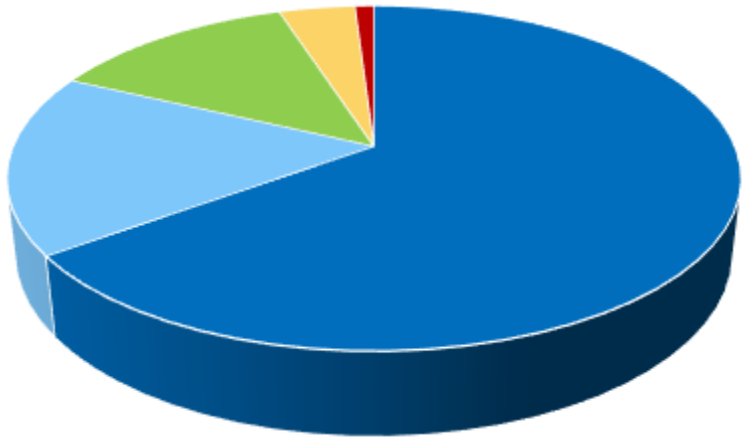
来源：金雅拓（Gemalto）2018年上半年全球数据泄露水平指数

- 仅在2018年上半年，全球就发生了945起较大型的数据泄露事件，其中社交媒体领域共发生了6次重大数据泄露事件，导致45亿条信息泄露（与2017年相比数量增加了133%）
- 平均每天有超过2,500万条记录被入侵或被泄露，包括医疗数据、信用卡数据和/或财务数据或个人身份信息

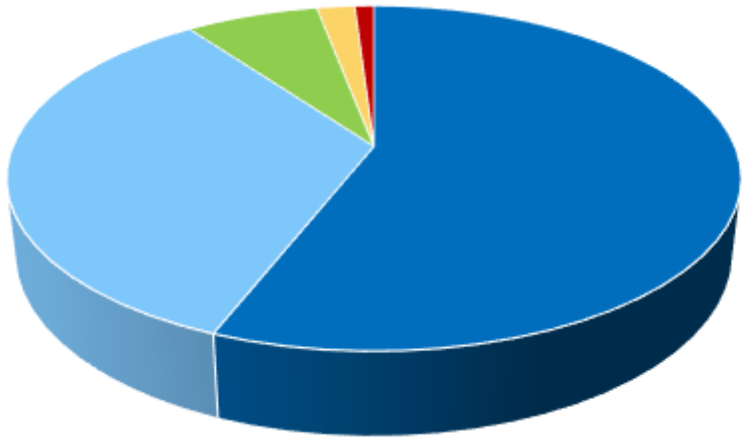
在945起数据泄露事件中

- **身份盗用**是最普遍的数据泄露类型
- **恶意外部人员攻击**是最主要的数据泄漏威胁源

Number of Breach Incidents by Type

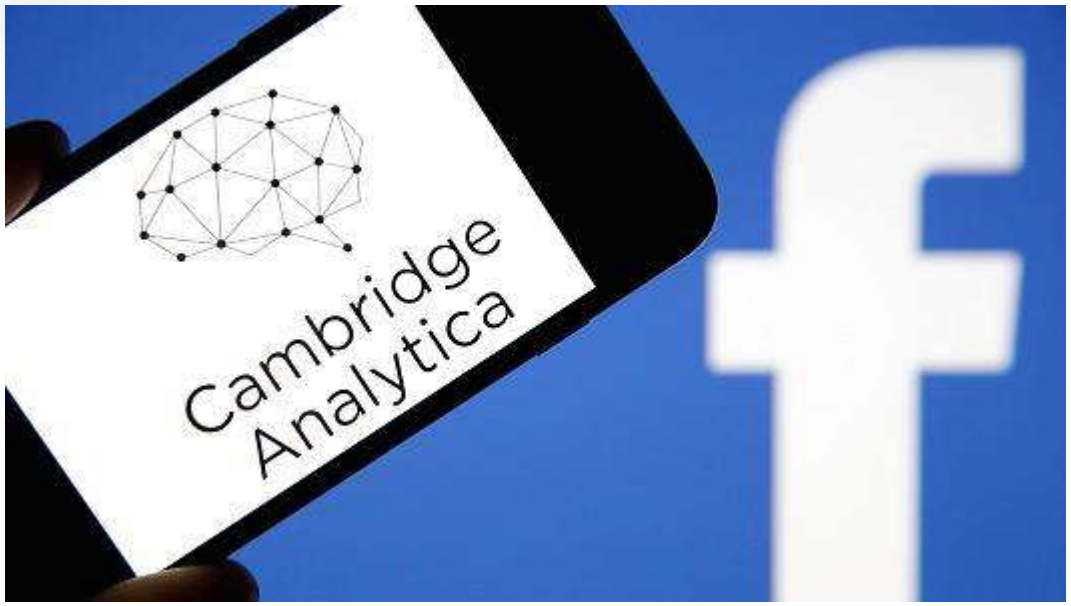


Number of Breach Incidents by Source



Identity Theft　Account Access　Financial Access　Malicious Outsider　Accidental Loss　Malicious Insider
Nuisance　Existential Data　Hackivist　Unknown

隐私泄露事件



Cambridge Analytica事件

- 2018年3月，美国纽约时报曝光了Facebook用户的个人信息在未经用户许可下被一家名为Cambridge Analytica的公司擅自使用的行为，媒体称这些数据被用来**向用户精准投放广告内容**，帮助2016年特朗普团队参选美国总统
- Facebook在公开回应中承认了有8700万用户私人信息被此公司进行了不正当使用
- 9月份，Facebook称黑客可利用漏洞获得3000万Facebook用户的账号信息
- 12月份，Facebook又宣布了另一个漏洞，第三方应用程序可以访问近700万用户未发布的照片

隐私泄露事件

Exactis事件

- 2018年6月，总部位于佛罗里达州的市场营销和数据聚合公司Exactis被发现将一个数据库放在可**公开访问的服务器上**
- 该数据库包含3.4亿条记录，涉及上亿美国成年人的个人信息和数百万公司的信息，信息总量超过2TB
- 库中所含个人信息非常详细，包括电话号码、家庭住址、电子邮箱地址和其他与个人特征高度相关的信息，例如个人爱好、习惯、年纪、性别等。

数据汇总公司Exactis

- 向世界各地的营销商和经销商收集和销售消费者和商业数据，收集了庞大的数据，**使公民，家庭和企业都可以轻松地画像、模仿或追踪**
- 由于数据库的公开，任何人都可以下载，**这次数据泄露使美国公民容易遭受黑客攻击**

隐私泄露事件

Under Armour事件

- 2018年3月，美国运动品牌Under Armour对外宣称，旗下健身应用MyFitnessPal因**存在数据漏洞**而遭到黑客攻击，多达1.5亿用户的数据被泄露

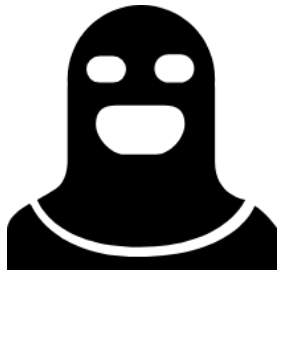
- 此次数据泄露事件涉及用户名、电子邮件地址和密码等账户信息

Twitter事件

- 2018年5月， Twitter 发布公告称，发现公司在存储用户密码时的一个 bug致使**部分用户的密码漏掉了加密这一过程**，每个访问密码日志文件的人，都能看到以明文存储的账号和密码，因此约有3.3亿 Twitter 用户密码存在泄漏风险

隐私泄露的危害

用户人身财产安全受到威胁



垃圾邮件

用户思想被操控

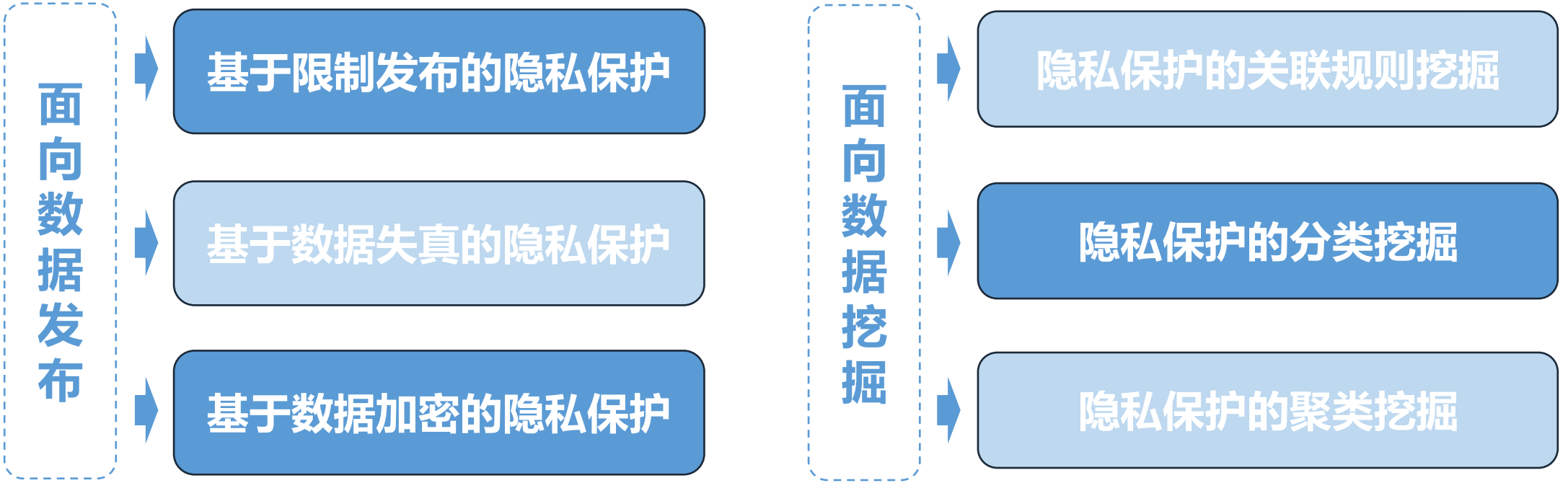
Cambridge Analytica精准投放广告，干预美国总统大选

威胁国家、企业安全

- 2016年，土耳其**国民信息数据库**遭到泄露，近5000万公民（包括现任总统）的身份证ID、姓名、父母亲姓名、性别、出生日期、居住地详细地址等信息泄露
- **筛选分析，精准犯罪**：公司微信群被克隆，员工汇给“老板”169万元

隐私保护技术初探

- 为了从大量的网络空间数据中获取有用信息，需要对其进行挖掘，在此过程中不免会造成数据隐私泄露，如何在获取有用信息的同时保护数据相关者的隐私变得尤为重要
- 数据隐私保护技术的研究主要分为两个方面：面向数据发布的隐私保护研究和面向数据挖掘的隐私保护研究



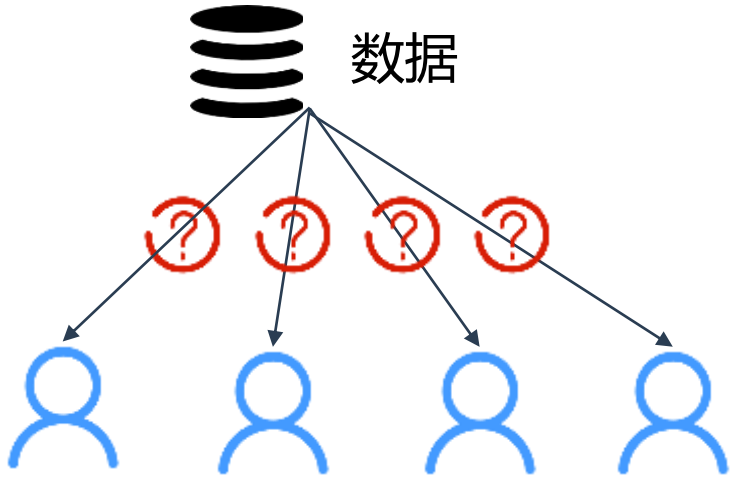
面向数据发布：基于限制发布的隐私保护

面向数据发布的隐私保护是在将数据公布给数据挖掘者之前，对数据进行扰动、加密、匿名等处理，将数据中的隐私藏起来

基于限制发布的隐私保护

有选择的发布原始数据、不发布或者发布精度较低的敏感数据

- 研究集中于数据匿名化
  - 研究更好的匿名化原则
  - 针对匿名化原则设计更高效的匿名化算法



隐藏用户身份和数据的对应关系

面向数据发布：基于数据失真的隐私保护

对原始数据进行扰动，目的是隐藏真实数据，只呈现出数据的统计学特征

基于数据失真的隐私保护技术主要包括随机化、阻塞、变形、交换等

失真后的数据满足

- 数据保持原本的某些特性不变
- 攻击者不能根据失真数据重构出真实的原始数据

- 随机化：在原始数据中加入随机噪声，从而隐藏真实的数据，保护敏感数据
- 数据交换：在记录之间交换数值以扰动真实数值，但同时要保留某些统计学特征

面向数据发布：基于数据加密的隐私保护

对原始数据进行加密，通过密码机制实现其他参与方对原始数据的不可见性以及数据的无损失性

由于加密技术可解决安全通信的问题，因此多应用于分布式应用

可使用的加密技术有

- 对称可加密搜索
- 安全多方计算
- 同态加密技术
- 数字信封技术
- Shamir秘密技术共享
- ……

可应用的分布式应用

- 分布式数据挖掘
- 分布式安全查询
- ……

两种数据存储模式

- 垂直划分：每个参与者只存储部分属性的数据，所有参与者存储的数据不重复
- 水平划分：将数据记录存储到多个参与者处，所有参与者存储的数据不重复

在两种存储模式中，每个参与者都只掌握了部分数据

面向数据发布的隐私保护总结

基于限制发布	基于数据失真	基于数据加密
有选择的发布原始数据、不发布或者发布精度较低的敏感数据	对原始数据进行扰动，目的是隐藏真实数据，只呈现出数据的统计学特征	对原始数据进行加密，通过密码机制实现其他参与方对原始数据的不可见性以及数据的无损失性

面向数据发布的隐私保护技术	优点	缺点
基于限制发布的隐私保护技术	发布的数据真实可靠	数据丢失部分信息
基于数据失真的隐私保护技术	算法效率较高	由于干扰使数据丢失部分信息
基于数据加密的隐私保护技术	数据的安全性和准确性均较高	计算开销很大

面向数据挖掘：关联规则的数据挖掘

关联规则是寻找在同一事件中出现的不同项目的相关性，关联规则挖掘是数据挖掘领域研究的重点之一，是从大量数据中挖掘数据项之间隐藏的关系，发现数据集中项集之间的关联和规则的过程，其通过置信度和支持度度量项集之间的规则

置信度：购买了一个商品之后又购买了另一种商品的可能性：

80/100=80%

支持度：购买关联商品的人数占总人数的比例：

80/1000=8%

设定最小置信度和最小支持度，当挖掘到的项集的置信度和支持度分别大于最小置信度和最小支持度时，就得到了关联规则

面向数据挖掘：关联规则的数据挖掘

关联规则数据挖掘应用之“购物篮分析”

- 商家通过对用户购物篮中的商品进行分析，研究用户的购买行为
- 研究结果对商家的决策起到了至关重要的作用

寻找商品之间隐藏的关联规则

关联规则的数据挖掘实例: 购物篮分析

在美国沃尔玛超市中，有一个有趣的现象，尿布和啤酒被摆在货架上一起出售，并且该措施使两种商品的销量双双增加，这是为什么呢?

- 沃尔玛为了能够准确了解顾客在其门店的购买习惯，利用其数据仓库系统对顾客的购物行为进行了购物篮分析，分析顾客经常一起购买的商品有哪些
- 结果发现，跟尿布一起购买最多的商品竟是啤酒!

隐藏的行为模式

在美国，一些年轻的父亲下班后经常要到超市去买婴儿尿布，而他们中有30%～40%的人同时随手为自己买一些啤酒

面向数据挖掘：隐私保护的关联规则挖掘

变换（distortion）

修改支持敏感规则的数据，使得规则的置信度和支持度小于一定的阈值而实现规则的隐藏

隐藏（blocking）

不修改数据，而是隐藏生成敏感规则的频繁项集，尽可能降低敏感规则的置信度或者支持度，以此使得需要保护或隐藏的规则不被挖掘出来

两类方法 ⚠ 都会影响对非敏感规则的挖掘

面向数据挖掘：隐私保护的分类和聚类挖掘

隐私保护的分类挖掘

分类：在数据集上构造分类函数或者分类模型，即分类器，将数据集中的数据项映射到给定的类别中，以用于类别的预测

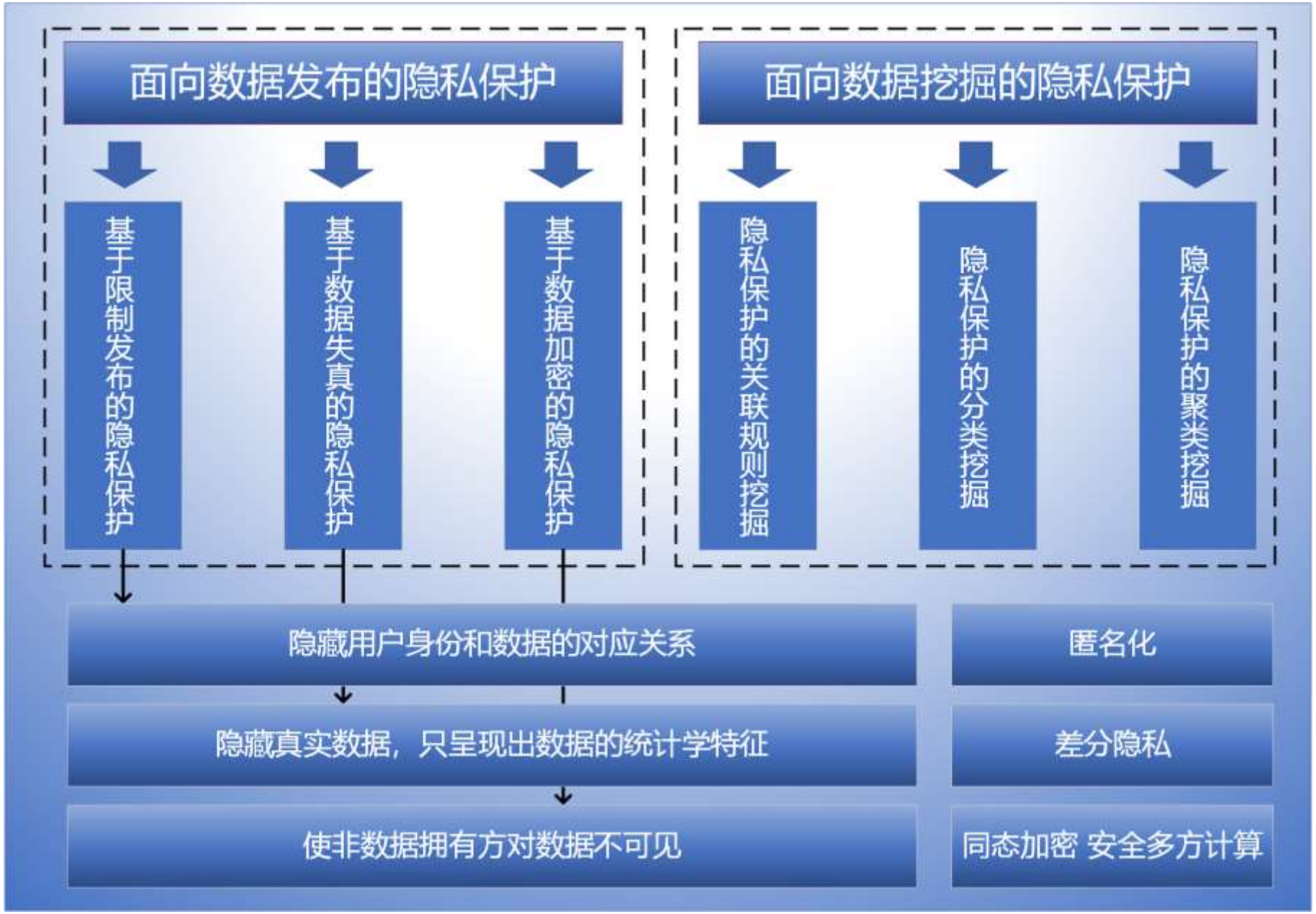
- 分类结果可能会暴露隐私信息
- 隐私保护的分类挖掘是指在数据挖掘的过程中，建立准确的、无隐私泄露的分类模型

隐私保护的聚类挖掘

聚类：将数据集中的数据根据相似性进行分类，最后的分类结果中同一个类别中的数据相似性越大越好，不同类别中的数据的相似性越小越好

- 与分类挖掘相同，由于聚类结果可能会暴露数据集中的隐私敏感信息，因此需要使用隐私保护技术保护敏感的分类结果信息

数据隐私保护技术总结


面向数据发布的隐私保护
面向数据挖掘的隐私保护
基于限制发布的隐私保护
基于数据失真的隐私保护
基于数据加密的隐私保护
隐私保护的关联规则挖掘
隐私保护的分类挖掘
隐私保护的聚类挖掘
隐藏用户身份和数据的对应关系
匿名化
隐藏真实数据，只呈现出数据的统计学特征
差分隐私
使非数据拥有方对数据不可见
同态加密 安全多方计算

第2节 匿名化

- 匿名化隐私保护模型
- 数据匿名化方法

如何安全的发布数据

病患记录表

姓名	性别	年龄	出生日期	邮政编码	家庭住址	疾病
张艳	女	36	1984-12-03	235023	海河市江流镇北路1号	流感
李磊	男	42	1978-06-25	235152	江宁市新源镇沙河路北	脂肪肝
王宇	男	35	1985-02-02	152030	丰宁市郭杜乡20号	糖尿病
赵静	女	34	1986-02-26	154263	江宁市新华镇南街2号	哮喘

- **若某医院想发布一个医疗数据集为其他机构提供研究信息，发布数据时该如何保护患者隐私？**

需要在确保发布的数据公开可用的前提下，保护用户的敏感数据和个人身份之间的对应关系

传统的匿名化方法

- 若想安全地发布医疗数据集供其他机构研究，需要**隐藏用户身份和数据的对应关系**
- 传统的匿名化方法
  - 删去表中容易关联到患者本人且研究价值不大的属性，即姓名和家庭住址
  - 或将姓名替换为假名

姓名	性别	年龄	出生日期	邮政编码	家庭住址	疾病
张艳	女	36	1984-12-03	235023	海河市江流镇北路1号	流感
李赛	男	42	1978-06-25	235152	江宁市新源镇沙河路北	脂肪肝
王宇	男	35	1985-02-02	152030	丰宁市郭杜乡20号	糖尿病
赵静	女	34	1986-02-26	154263	江宁市新华镇南街2号	哮喘

性别	年龄	出生日期	邮政编码	疾病
女	36	1984-12-03	235023	流感
男	42	1978-06-25	235152	脂肪肝
男	35	1985-02-02	152030	糖尿病
女	34	1986-02-26	154263	哮喘

使用传统的匿名化方法处理数据能否保护患者隐私？

数据表中涉及的概念

姓名	性别	年龄	出生日期	邮政编码	家庭住址	疾病
张艳	女	36	1984-12-03	235023	海河市江流镇北路1号	流感
李磊	男	42	1978-06-25	235152	江宁市新源镇沙河路北	脂肪肝
王宇	男	35	1985-02-02	152030	丰宁市郭杜乡20号	糖尿病
赵静	女	34	1986-02-26	154263	江宁市新华镇南街2号	哮喘

准标识符 （Quasi-Identifiers, QID）是与其他数据表进行链接以标识个体身份的属性或属性组合，如性别 、出生日期、邮政编码等，其选择**取决于进行链接的外部数据表**

标识符（Identifiers）是唯一标识个体身份的属性或者属性的集合，如姓名、身份证号等

敏感属性（Sensitive Attributes, SA)是发布时需要保密的属性，如薪资、健康状况等

链接攻击

攻击者掌握的选民信息表

姓名	性别	年龄	出生日期	邮政编码	登记日期
钱晶	女	24	1996-05-02	256230	2020-07
王宇	男	35	1985-02-02	152030	2020-07
周平	男	32	1988-12-21	152630	2020-07
秦桦	女	29	1991-03-16	152410	2020-07
吴沛	男	31	1989-08-30	256230	2020-07

匿名后的医疗数据集

性别	年龄	出生日期	邮政编码	疾病
女	36	1984-12-03	235023	流感
男	42	1978-06-25	235152	脂肪肝
男	35	1985-02-02	152030	糖尿病
女	34	1986-02-26	154263	哮喘

- 传统的匿名化方法无法抵抗**链接攻击**

攻击者通过对发布数据和其他渠道获取的数据进行链接操作推理出隐私信息

攻击者将医疗数据集与其他公共数据集的**准标识符**联系起来，推断王宇患有糖尿病

链接攻击

- 2000年，卡内基梅隆大学教授Latanya Sweeney在《简单的人口统计往往能识别出人的独特性》报告中提到在基于美国选举人公共注册信息的基础上
  - 87%的美国人基于邮编、性别、出生日期即有可能被识别出个人身份
  - 53%的美国人基于地址、性别、出生日期即有可能被识别出个人身份
  - 18%的美国人基于县、性别、出生日期即有可能被识别出个人身份


Simple Demographics Often Identify People Uniquely

- **包括上述个人信息的数据字段的公开可能会导致隐私泄露**
- 该研究曾使用麻省总医院的出院数据和选举投票的注册数据进行匹配，最终链接出某麻省议员的住院信息

美国在线隐私泄漏事件

2006年，为了学术研究，美国在线（AOL）公开了65万名用户在3个月内的2000万次搜索请求

虽然**AOL用户名都被随机ID代替**，但是如果对某一用户的搜索记录进行分析，有可能判断出其身份和行为

提供了439MB的压缩包，甚至还有一个详尽的Readme文件

《纽约时报》发现匿名ID为4417749的用户为来自佐治亚州利本市的62岁寡妇Thelma Arnold

简单地删除敏感字段或者将姓名替换为假名不能保证个人隐私的安全

匿名化隐私保护模型：k-anonymity

k-anonymity是由Pierangela Samarati和Latanya Sweeney于1998年提出的隐私保护模型，有效解决了链接攻击问题

- 等价类：匿名化后的数据表中具有相同准标识符的若干记录称为一个等价类
- k-anonymity：将k个记录放入一个等价类中，**要求任意一条记录与其他至少k-1条记录相似而不可区分**，这样数据中的每一条记录都能找到与之相似的记录，降低了数据的识别度

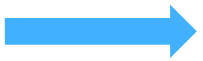
如果一条记录由于样本太少而无法找到k-1条相似的记录，那么这条数据不应当被纳入数据集

k-anonymity

攻击者进行链接攻击时，由于对任意一条记录的攻击，都会同时关联到等价类中的其他k-1条记录，因此攻击者无法确定特定用户

公布的病患数据

年龄	邮政编码	疾病
52	123023	心脏病
32	120156	糖尿病
59	123152	心脏病
30	120162	糖尿病
56	123485	心脏病
35	120154	哮喘



病患数据的3-anonymity版

年龄	邮政编码	疾病
5*	123***	心脏病
5*	123***	心脏病
5*	123***	心脏病
3*	1201**	糖尿病
3*	1201**	糖尿病
3*	1201**	哮喘

同质性攻击

k-anonymity保证了单独个体被准确标识的概率最大为1/k，但却无法保证隐私不被泄露

病患数据的3-anonymity版本

年龄	邮政编码	疾病
5*	123***	心脏病
5*	123***	心脏病
5*	123***	心脏病
3*	1201**	糖尿病
3*	1201**	糖尿病
3*	1201**	哮喘

同质性攻击：在数据匿名化过程中，由于没有对敏感属性进行约束，最终结果可能会造成隐私泄露

如果一名选民的年龄和邮政编码符合第一个等价类要求，那么攻击者可**推断该选民可能患有心脏病**

背景知识攻击

k-anonymity保证了单独个体被准确标识的概率最大为1/k，但却无法保证隐私不被泄露

病患数据的3-anonymity版本

年龄	邮政编码	疾病
5*	123***	心脏病
5*	123***	心脏病
5*	123***	心脏病
3*	1201**	糖尿病
3*	1201**	糖尿病
3*	1201**	哮喘

背景知识攻击：攻击者可以通过掌握的足够的相关背景知识以很高的概率确定敏感数据与个体的对应关系，得到隐私信息

如果一名选民的年龄和邮政编码符合第二个等价类要求，并且攻击者发现他不像是患有哮喘，那么攻击者可**推断该选民可能患有糖尿病**

l-diversity

l-diversity在k-anonymity的基础上，要求保证**每一个等价类的敏感属性至少有l个不同的值**，即每个用户的敏感属性值在等价类中可以找到与此值不同的至少l-1个属性值，使攻击者最多只能以1/l的概率确认某个用户的敏感信息

病患数据的3-anonymity版本

年龄	邮政编码	疾病	
5*	123***	心脏病	1种取值
5*	123***	心脏病	
5*	123***	心脏病	
3*	1201**	糖尿病	2种取值
3*	1201**	糖尿病	
3*	1201**	哮喘	

3-diversity举例（敏感属性至少三种取值）

Data table of *l*-diversity, where *l*=3

	Non-Sensitive			Sensitive
	Zip Code	Age	Nationality	Condition
1	1305*	<=40	*	Heart Disease
4	1305*	<=40	*	Viral Infection
9	1305*	<=40	*	Cancer
10	1305*	<=40	*	Cancer
5	1485*	>40	*	Cancer
6	1485*	>40	*	Heart Disease
7	1485*	>40	*	Viral Infection
8	1485*	>40	*	Viral Infection
2	1306*	<=40	*	Heart Disease
3	1306*	<=40	*	Viral Infection
11	1306*	<=40	*	Cancer
12	1306*	<=40	*	Cancer

l-diversity

l-diversity保证攻击者最多只能以1/l的概率确认某个用户的敏感信息，但无法保证隐私不被泄露

等价类中敏感值的分布与整个数据集中敏感值的分布具有明显的差别，攻击者可以以一定概率猜测目标用户的敏感属性值

				某疾病检测结果
				阴性
				阳性
				阴性
				…

1000条记录中有1%的阳性记录和99%的阴性记录，阳性检测结果更为敏感

2-diversity

- 每个等价类中疾病检测结果必须包含阴性和阳性两种结果
- 假设某等价类中有一半阳性记录和一半阴性记录，相比于整体1%阳性的概率，该等价类中的个体都有1/2的概率被认为是阳性，具有严重的隐私风险

l-diversity

l-diversity保证攻击者最多只能以1/l的概率确认某个用户的敏感信息，但无法保证隐私不被泄露

l-diversity并没有考虑语义信息也会为隐私信息带来泄露的风险

				工资
				5000
				5500
				6000
				…

敏感属性为工资

l-diversity

- 若某一个等价类中的工资这一属性的属性值全部在一个固定区间内，那么攻击者并不需要知道详细的属性值就可以通过这个区间就可以判断用户的工资水平

t-closeness

在k-anonymity和l-diversity的基础上，t-closeness考虑了敏感属性的分布问题，要求**所有等价类中的敏感属性的分布尽量接近该敏感属性的全局分布**，差异不能超过阈值t

				工资
				5000
				5500
				6000
				…

敏感属性为工资

t-closeness

- 保证工资的分布和整体的分布类似，进而很难推断出某人工资的高低

k-anonymity，l-diversity 和 t-closeness**以信息损失为代价**，隐私保护效果逐个提高，但是它们一定能保证隐私不被泄露吗

隐私泄露风险

k-anonymity、l-diversity和t-closeness不能够完全保护隐私不被泄露

- 造成较大的信息损失，信息损失可能会使数据使用者们做出误判
- 对所有敏感属性提供了相同程度的保护并且没有考虑语义关系，造成了不必要的信息损失

针对不同的问题，提出不同的匿名技术

- 不同的用户对于隐私信息有着程度不同的隐私保护要求
- 属性与属性之间的重要程度并不相同
- 没有考虑数据动态更新后重发布的隐私保护问题

个性化匿名技术

带权重的匿名策略

动态数据匿名化

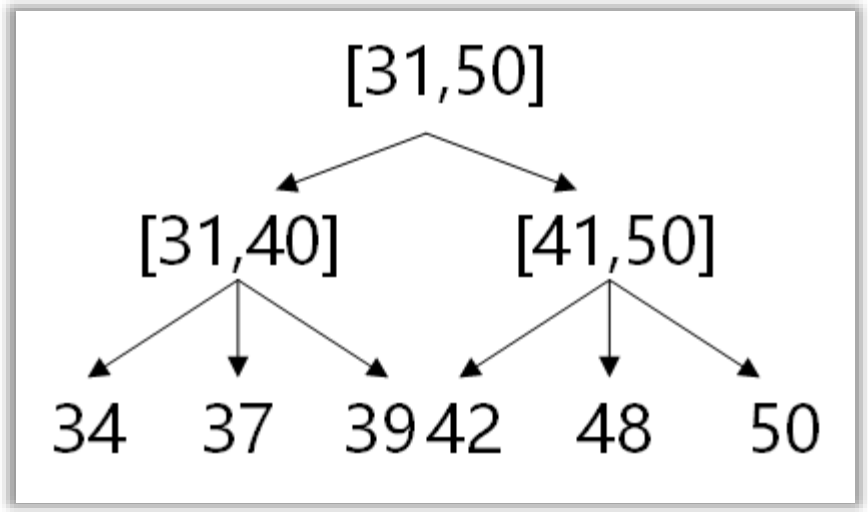
数据匿名化方法

- 实现匿名化的方法有泛化、抑制、聚类、微聚集、分解、置换等
- 目前提出的匿名化主要通过**泛化和抑制**实现，它们能保持发布前后数据的真实性和一致性

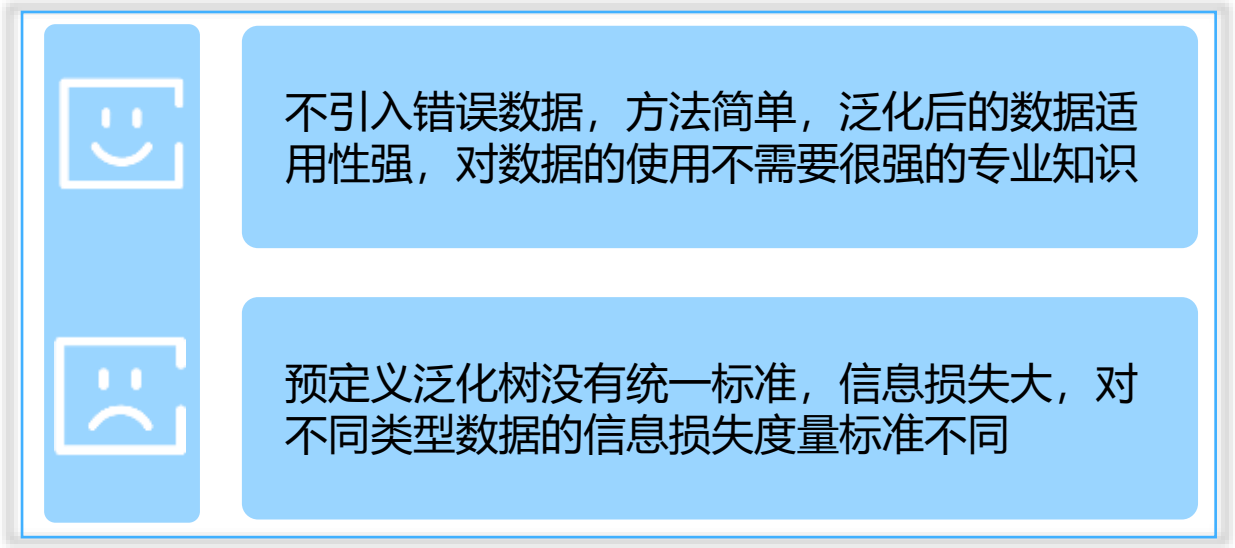
匿名化方法	思想
泛化	用更抽象、概括的值或区间代替精确值
抑制	将数据表中的数据直接删除或隐藏
聚类	按照给定的规则将数据集分成各类簇，尽量保证簇内对象相似，不同簇的对象相异
微聚集	相似的数据划分在同一个类中，每个类至少有k条记录，用类质心代替类中所有记录的准标识符属性值
分解	根据敏感属性值对数据表分组，尽量使得同一组的敏感属性值不同，将分组后的数据表拆分为分别包含准标识符属性信息和包含敏感属性信息的两张表
置换	对数据表分组，把每组内的敏感属性值随机交换，打乱顺序，再拆分数据表，对外发布

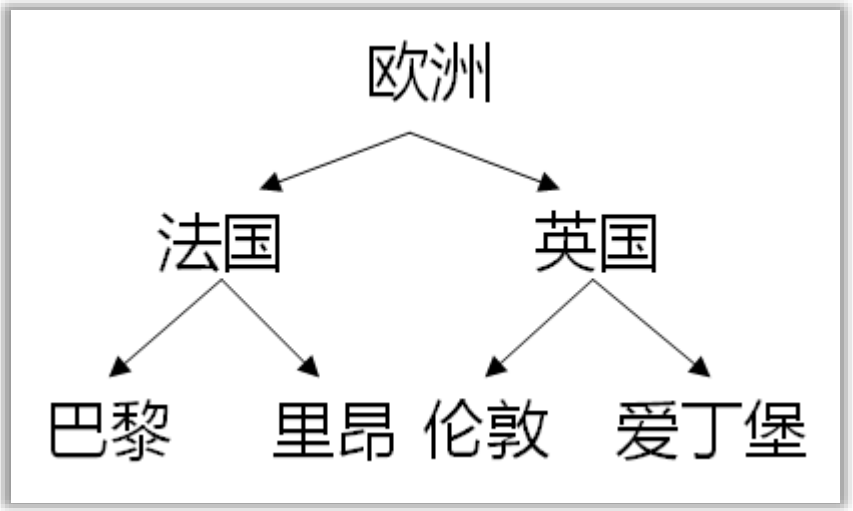
泛化

- **将准标识符的属性用更一般的值或者区间代替**
- 准标识符属性值有数值型和分类型
  - 数值型：值被一个覆盖精确数值的区间代替
  - 分类型：用一个更一般的值代替原值



数值型数据泛化树





分类型数据泛化树

泛化树

顶层的取值最模糊

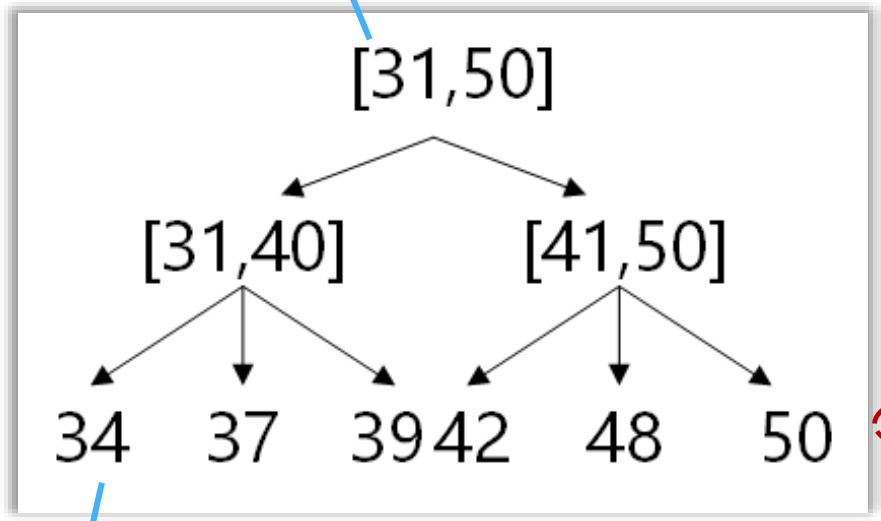
泛化域2：{法国，英国}

泛化域3：{欧洲}

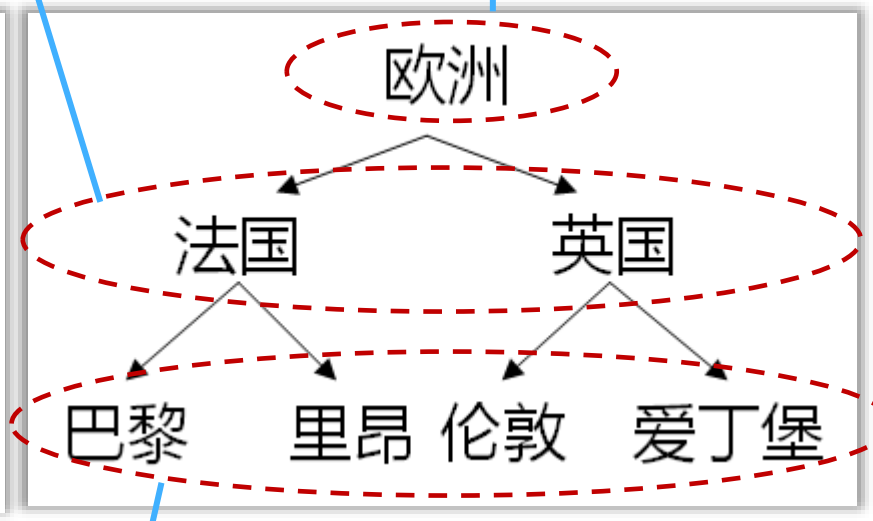
取值模糊

底层的取值可以泛化为高层的取值，树中每层的取值构成一个泛化域

取值具体



数值型数据泛化树



分类型数据泛化树

底层的取值最具体

泛化域1：{巴黎，里昂，伦敦，爱丁堡}

域泛化（全局泛化）

将一个给定的属性域泛化为一般域，将准标识符属性值从底层开始**同时向上泛化**，一层层泛化直到满足隐私保护要求，然后**同时停止泛化**

信息损失太大

全域泛化	子树泛化	兄弟节点泛化
某个属性的全部值必须在同一层上进行泛化	泛化树中的同一个父亲节点下的所有孩子节点全部泛化或者全部不泛化	在同一个父亲节点下，如果对部分孩子节点进行泛化，其他兄弟节点不要求泛化，父亲节点只能代替泛化了的孩子节点

值泛化（局部泛化）

将原始属性域中的每个值直接泛化成一般域中的唯一值，将准标识符属性值从底层向上泛化，但是**可以泛化到不同的层次**

单元泛化

对某个属性的一部分值进行泛化，另一部分值保持不变

多维泛化

对多个属性的值同时泛化，只需要对不符合限制要求的等价类进行泛化，要求一个等价类中所有记录都泛化成相同的值

抑制

抑制，又称隐藏、隐匿，是将准标识符属性值从数据集中**直接删除或者用诸如“*”之类的代表不确定值的符号来代替**，与泛化结合使用

记录抑制	值抑制	单元抑制
对数据表中的某条记录进行抑制处理	对数据表中的某个属性的值全部进行抑制处理	对数据表中某个属性的部分值进行抑制处理

第3节 差分隐私

- 差分隐私基础
- 数值型差分隐私
- 非数值型差分隐私

匿名化与差分隐私

不能提供足够的数学保障，没有严格定义攻击模型，对攻击者所具有的知识未能作出定量化定义，无法抵抗背景知识攻击

不能提供严格和科学的方法证明其隐私保护水平，当模型参数改变时，无法对其隐私保护水平进行定量分析

匿名化

差分隐私

假设攻击者掌握最大知识背景，即能够获得除目标记录外所有其它记录的信息

对隐私保护进行了严格的定义并提供了量化评估方法，使不同参数处理下的数据集所提供的隐私保护水平具有可比较性

与匿名化相比，差分隐私是一种严格的可证明的隐私保护模型

差分攻击

姓名	是否患病
张三	0（不患病）
李四	1
王五	1
钱六	1

$f(3)$

2

$f(4)$

3

计数查询服务： $f(i) = count(i)$
查询数据集中前i行患病的记录数量

第四行代表的用户患病了！
若已知该用户是钱六，则可**推断钱六患病**

为抵抗差分攻击，差分隐私要求保证任意一个个体在数据集中或者不在数据集中时，对最终发布的查询结果几乎没有影响

差分隐私思想

只有一条数据不同的两个数据集

姓名	是否患病
张三	0
李四	1
王五	**1**
…	…

姓名	是否患病
张三	0
李四	1
钱六	**1**
…	…

随机算法$M(D1)$ → 查询结果是100的概率是99%

随机算法$M(D2)$ → 查询结果是100的概率是98%

查询有多少人患病，用随机算法对信息做扰动

$$\frac{99\%}{98\%} = 1.01$$

同一查询在两个数据集上产生相同结果的概率的比值接近于1

对数据集中的每个个体的隐私进行保护

概念介绍

隐私保护机制

对数据集D的各种映射函数被定义为查询（Query），用$F = \{f1, f2, \cdots\}$来表示一组查询，算法M对查询 F 的结果进行处理，使之满足隐私保护的条件，此过程称为隐私保护机制

邻近数据集

设数据集D 和D′具有相同的属性结构，两者的对称差记作$D\Delta D'$，$|D\Delta D'|$表示$D\Delta D'$中记录的数量，**若$|D\Delta D'| = 1$，则称D和D′为邻近数据集**（Adjacent Dataset)

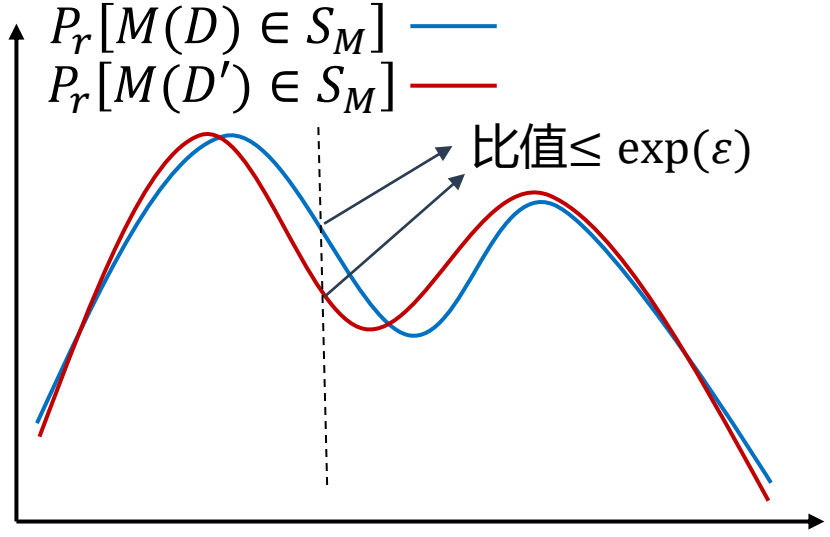
例如，设$D = \{1,2,3,4,5\}$，$D' = \{1,2,4\}$，则$D\Delta D' = \{3,5\}$，$|D\Delta D'| = 2$

差分隐私定义

差分隐私（Differential Privacy，DP）的定义：设有一个随机算法M，P_M为算法M所有可能的输出构成的集合，如果对于任意两个邻近数据集D和D'以及P_M的任意子集S_M，算法M满足：

$$P_r[M(D) \in S_M] \leq \exp(\varepsilon) \times P_r[M(D') \in S_M]$$

则称算法M提供ε –差分隐私保护，其中**参数ε称为隐私保护预算**



- $\varepsilon = 0$：攻击者无法区分相邻数据集，保护程度最高，数据可用性最差
- ε增大：保护程度越来越低，ε过大，则会造成隐私泄露
- ε通常取很小的值，例如0.01，0.1，或者$\ln 2$，$\ln 3$等

ε的取值应当结合具体需求设定以达到输出结果的安全性与可用性的平衡

抵抗差分攻击

姓名	是否患病
张三	0（不患病）
李四	1
王五	1
钱六	1

$f(3)$

?

$f(4)$

?

两次查询返回的结果以几乎相同的概率来自集合$\{1.5，2，2.5\}$

计数查询服务：$f(i) = count(i) + noise(i)$
查询数据集中前i行患病的记录数量
$f(i)$是提供ε −差分隐私保护的查询函数

针对统计输出的随机化方式使攻击者无法得到查询结果间的差异，保证了数据集中每个个体的安全

差分隐私的实现

差分隐私可以通过在查询函数的返回值中加入噪声来实现

增大加入的噪声 ⇨ 数据可用性下降

完全随机的数据没有意义

减小加入的噪声 ⇨ 数据安全性下降

无法保护个体隐私

全局敏感度

全局敏感度：设有函数$f: D \to R_d$，输入为数据集，输出为d维实数向量。对**任意的邻近数据集D 和D′**，

$$GS_f = \max_{D,D'} \|f(D) - f(D')\|_1$$

称为函数f的全局敏感度，其中$\|f(D) - f(D')\|_1$是$f(D)$和$f(D')$之间的1 –阶范数距离

所有维度上的距离之和，若查询结果是一维的(select a from ...)，距离为$|a - a'|$（两个数字的差的绝对值），若是二维的（select a, b from ...），距离为$|a - a'| + |b - b'|$

姓名	是否患病
张三	0
李四	1
王五	1
钱六	1

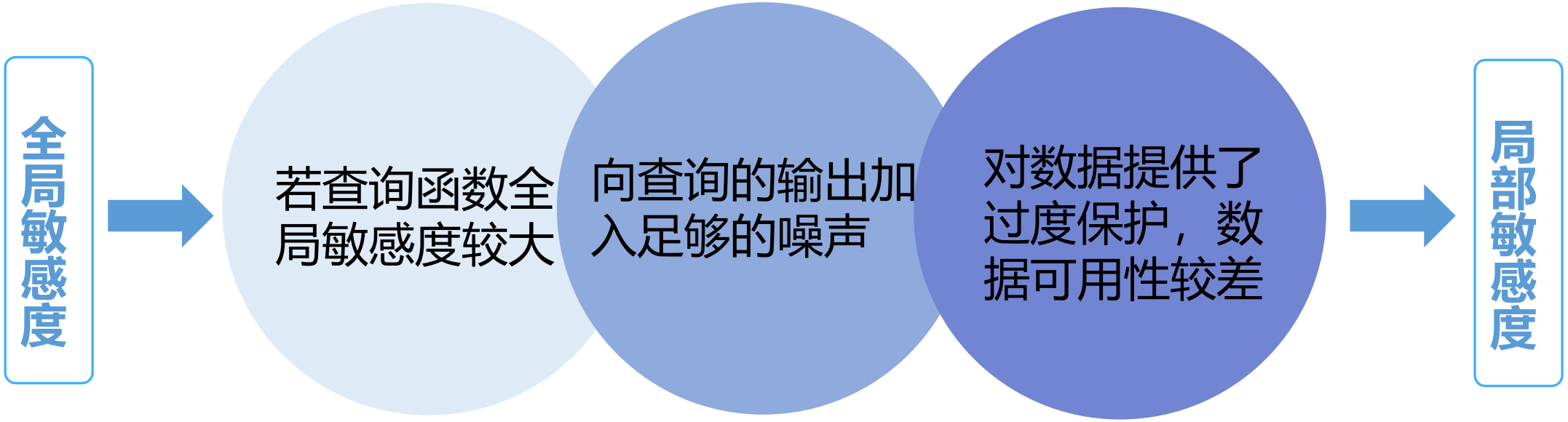
姓名	是否患病
张三	0
李四	1
钱六	1

以计数查询函数f（查询患病人数）为例，对任意的D和D'，由于一条记录的有无只会使f输出的差值为1，因此该函数的全局敏感度为1

数据集D：$f(D)$ =3　　数据集D'：$f(D')$=2

全局敏感度反映了一个查询函数在一对邻近数据集上进行查询时变化的最大范围，它与数据集无关，由查询函数本身决定

从全局敏感度到局部敏感度



局部敏感度

局部敏感度：设有函数$f: D \to R_d$，输入为数据集，输出为d维实数向量。对于**给定的数据集D和它的任意邻近数据集D′**，

$$LS_f = \max_{D'} \|f(D) - f(D')\|_1$$

称为函数f在D上的局部敏感度

全局敏感度

- 对任意的邻近数据集D 和D′
- 只由查询函数决定

给定的数据集与全局敏感度中使1-阶范式距离达到最大的数据集相同时，局部敏感度就等于全局敏感度

局部敏感度

- 对给定的数据集D 和它的任意邻近数据集D′
- 由查询函数和给定的数据集中的数据共同决定

数值型差分隐私：拉普拉斯和高斯机制

数值型差分隐私的实现机制有**拉普拉斯机制**和高斯机制，通过在查询结果中加入随机噪声实现隐私保护

拉普拉斯机制提供的是严格的$(\varepsilon, 0)$ −差分隐私保护，而高斯机制提供的是松弛的(ε, δ) −差分隐私保护

拉普拉斯分布

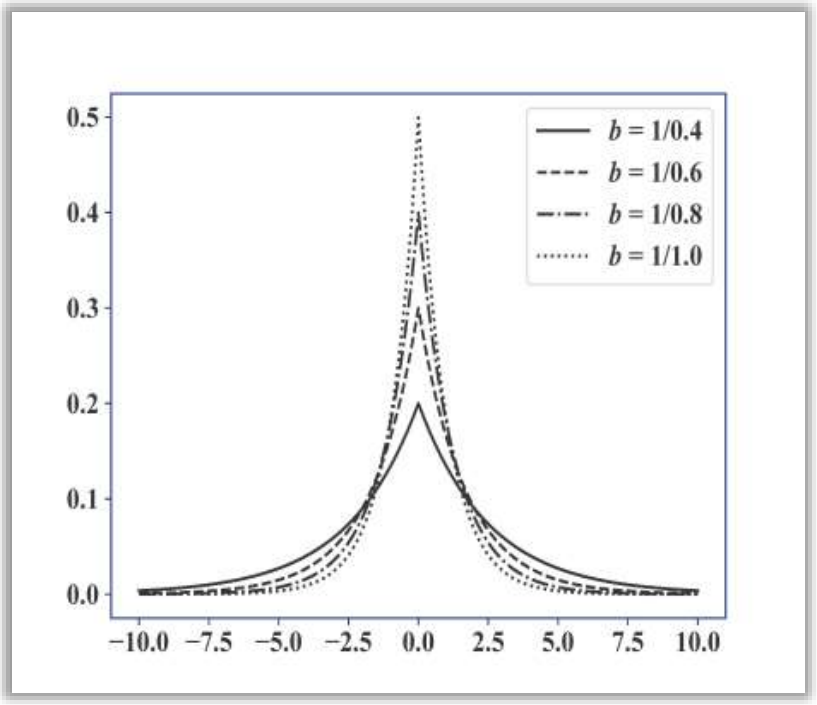
拉普拉斯分布是一种连续的概率分布，其概率密度函数为：

$$f(x|\mu, b) = \frac{1}{2b} exp\left\{-\frac{|x-\mu|}{b}\right\}$$

其中位置参数为μ，尺度参数为$b(b > 0)$，该分布的期望值为μ，方差为$2b^2$

记位置参数μ为0，尺度参数为b的拉普拉斯分布为Lap(b)，它的概率密度函数为：

$$p(x) = \frac{1}{2b} exp\left\{-\frac{|x|}{b}\right\}$$



不同尺度参数下的拉普拉斯分布图像

拉普拉斯机制的定义

拉普拉斯机制是一种广泛应用于数值型差分隐私的隐私保护机制，其思想为在数值型数据的查询结果中添加随机的满足拉普拉斯分布的噪声来实现差分隐私保护

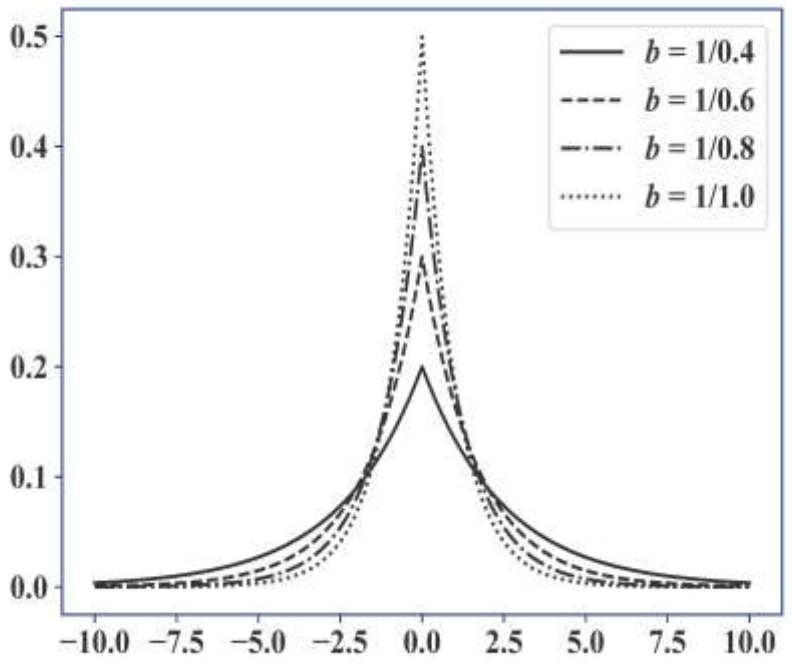
对于任意的数据集D和函数$f: D \to R^d$，其全局敏感度为Δf，若随机算法M的输出结果满足

$$M(D) = f(D) + Lap(\frac{\Delta f}{\varepsilon})$$

则算法M满足$(\varepsilon, 0)$ −差分隐私保护，其中$Lap(\frac{\Delta f}{\varepsilon})$为添加的随机噪声，服从尺度参数为$b = \frac{\Delta f}{\varepsilon}$的拉普拉斯分布

服从拉普拉斯分布的随机噪声

$Lap(\frac{\Delta f}{\varepsilon})$ 为服从尺度参数为$b = \frac{\Delta f}{\varepsilon}$的拉普拉斯分布的随机噪声



不同尺度参数下的拉普拉斯分布图像

设全局敏感度Δf =1

- 噪声量与Δf成正比，与ε成反比
- 随着ε的减小，对输出的混淆程度增强（使真实值变为一个和真实值具有较大差别的值的概率越大），保护程度越高
- Δf越大，加入的噪声越大，保护程度越高，但是当Δf较大时，往往会对数据提供过度的保护

低隐私保护 → 高隐私保护

拉普拉斯机制的应用

统计查询

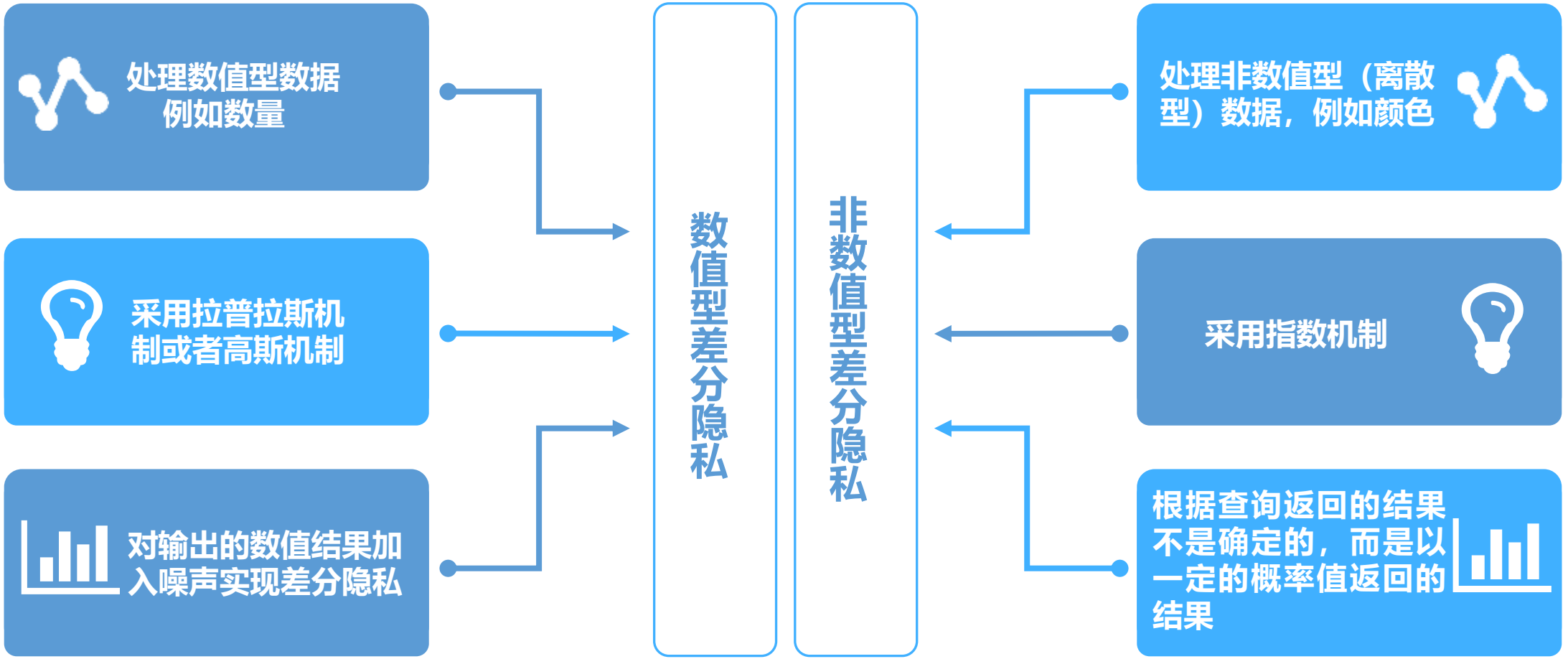
查询数据集中有多少项数据满足给定的条件，每一条数据可能满足也可能不满足，因此其全局敏感度为1，可直接在查询结果中加上$Lap(\frac{1}{\varepsilon})$ 实现差分隐私保护

直方图查询

数据直方图中的数据表示每一个单元有多少条记录，一个具有k个单元的直方图可以看做k个单独的计数查询，由于一行数据的增加或删除只会对这行数据对应的那个单元的计数造成影响，因此每个单元之间是独立的，其敏感度也为1，可直接在查询结果中加上$Lap(\frac{1}{\varepsilon})$ 实现差分隐私保护

- 例如，从有1000个名字的列表（人口普查参与记录）中查询参与者最常用的名字
- 每个人只能有一个名字，使用每个名字的人数是一个直方图单元格，任务是产生“最佳”答案，即为人数最多的直方图单元格

非数值型差分隐私：指数机制



指数机制的定义

在数据集D上进行查询时，对于非数值型数据，我们将可能得到的查询结果的集合称为输出域$Range$，域中的每一个值为实体对象r，以查询**得病人数最多的疾病**为例，当D中有四种疾病时，输出域为四种疾病的集合，每一种疾病是一个实体对象

函数$q(D,r)$为r的可用性函数，用来评估输出的r的优劣程度，以查询**得病人数最多的疾病**为例，可用性函数为计算某种疾病的得病人数

设随机算法M输入为数据集D，输出为一实体对象$r \in Range$，$q(D,r)$为可用性函数，Δq为函数$q(D,r)$的敏感度，若算法M以正比于$exp\left\{\frac{\varepsilon q(D,r)}{2\Delta q}\right\}$的概率从$Range$中选择并输出$r$，那么算法$M$提供$\varepsilon$-差分隐私保护

$\Delta q = \max_{D,D'} \|q(D,r) - q(D',r)\|_1$，当可用性函数为计算某种疾病的得病人数时，由于一条记录的有无造成的可用性函数输出的最大变化值为1，因此其敏感度为1

为每个r计算$exp\left\{\frac{\varepsilon q(D,r)}{2\Delta q}\right\}$，对所有结果进行归一化，由此确定每个r输出的概率值

指数机制的应用

假设查 计算某种疾病的得病人数，可
化

$\varepsilon = 0$时，各选项被输出的概率相同，无法根据邻近数据集的查询结果判断单条记录的属性值，所以隐私保护程度最高，但数据失去可用性

指数机制下对得病 病的查询

疾病	可用性 $q(D,r)$ $(\Delta q = 1)$	概率		
		$\varepsilon = 0$	$\varepsilon = 0.1$	$\varepsilon = 1$
流感	30	0.25	0.424	0.924
脂肪肝	25	0.25	0.330	0.075
哮喘	8	0.25	0.141	1.5E-05
糖尿病	2	0.25	0.105	7.7E-07

ε较小时，各选项在可用性上的差异被抑制，被输出的概率趋于相同

ε增大，可用性最好的选项被输出的概率不断增大

ε的选择，需要在保障数据可用的前提下进行权衡

第4节 同态加密

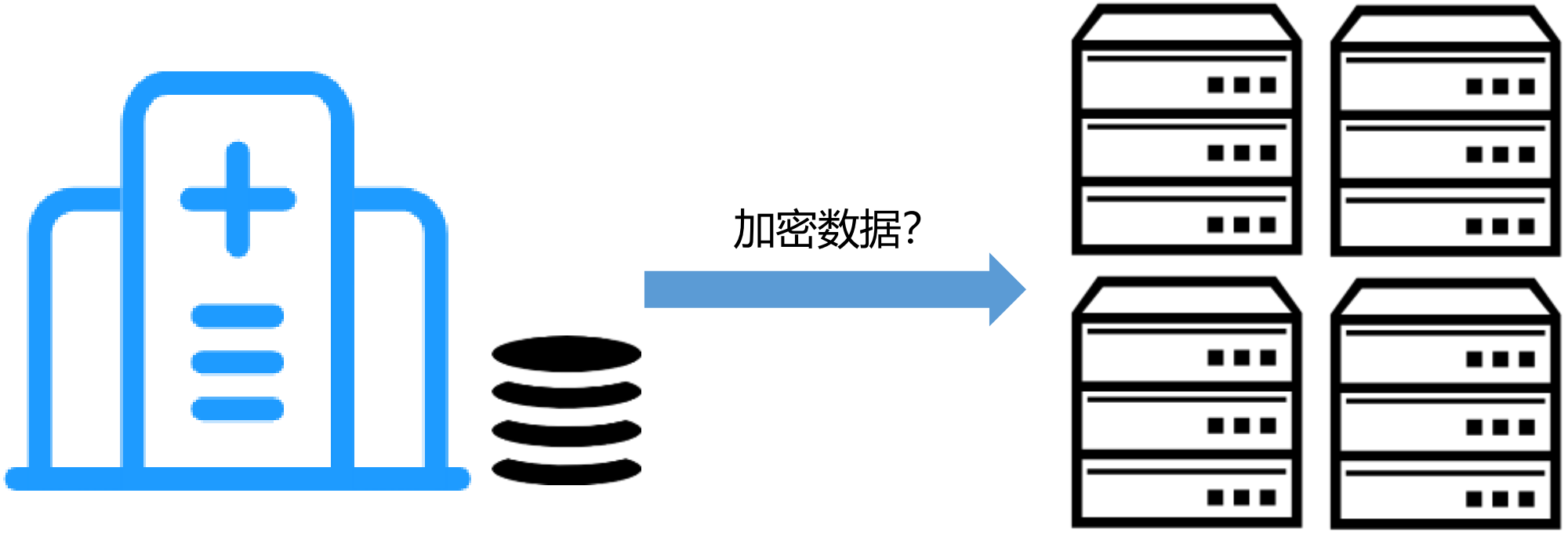
- 同态加密基础
- 半同态加密
- 全同态加密

同态加密的场景：安全的数据外包

医院拥有数据却无计算能力

数据计算平台拥有计算能力却不一定可信

数据加密实现安全数据外包



加密是一种众所周知的保护敏感信息隐私的技术，但是数据在加密之后如果想要对数据进行运算，就必须要对密文进行解密，增加了不安全因素

同态加密的提出



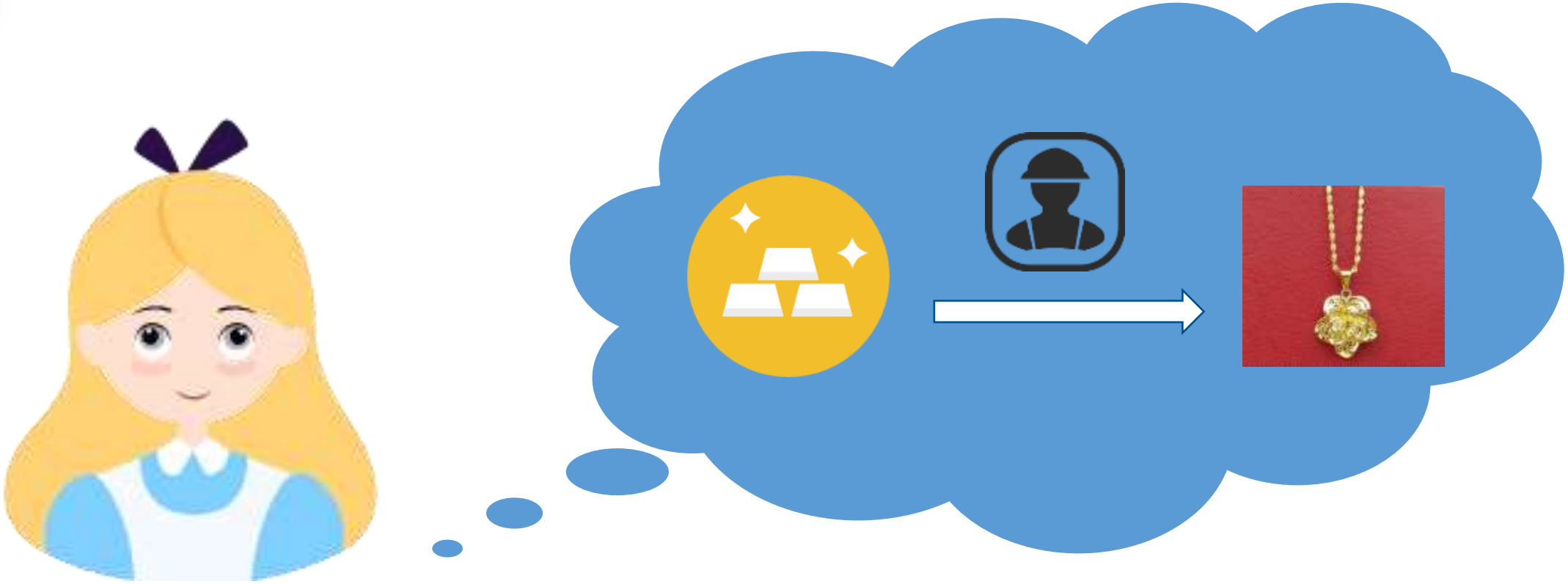
- 1978年，Ronald L. Rivest，Leonard Adleman 和 Michael L. Dertouzos 以银行为应用背景提出了同态加密（Homomorphic Cryptosystem, HC）的概念
- Ronald L. Rivest, Leonard Adleman分别是RSA算法（1977年提出）提出者中的R和A
- RSA算法可以实现乘法同态

A way to delegate processing of your data, without giving a way access to it.

— Craig Gentry（第一个全同态加密构造者）

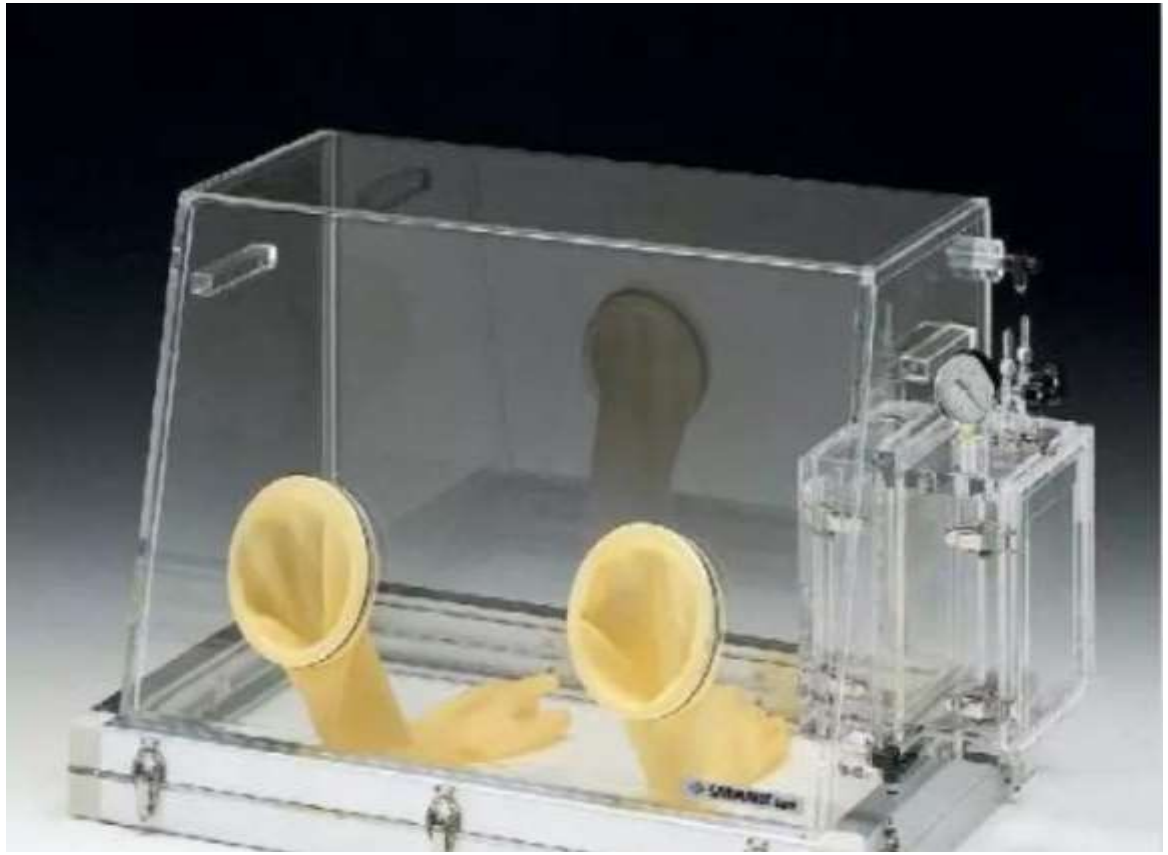
思想：对密文直接进行操作，且计算结果的解密值与对应明文的计算结果相同

同态加密思想



- Alice有一块金子，她想请工人把金子打造成一条项链，该如何防止金子被盗
- 想出一种办法，让工人可以对金子进行加工，但是却不能拿走任何金子

同态加密思想



Alice可以这样做

- 把金子锁在一个密闭的箱子里，为箱子安装一副手套
- 工人通过手套对箱子内部的金子加工，但是由于箱子是锁着的，因此工人无法拿走金子
- 工人加工完后，Alice用钥匙将箱子打开，得到了加工好的项链

一种不需要拿到金子本身就可以加工金子的方法

同态加密思想

用户密钥

加密算法

- 把金子**锁**在一个密闭的**箱子**里，为箱子安装一副手套
- 工人通过手套对箱子内部的金子加工，但是由于箱子是锁着的，因此工人无法拿走金子
- 工人加工完后，Alice用钥匙将箱子打开，得到了加工好的项链

将数据用同态加密方式进行加密

在不知道原数据的情况下对加密后的数据进行操作

对结果进行解密，得到处理后的结果

同态加密算法

Encrypt算法
用公钥将明文加密为密文

Decrypt算法
用私钥将密文解密为明文

KeyGen算法
由安全参数计算一对公私钥

Evaluate算法
在密文上进行运算

同态加密算法的四个组成部分

同态加密的发展

1978年Ronald Rivest等人提出同态加密的概念

国内外经过40年的研究，不断提出新的加密方案，并逐渐应用于实际中

2009年，Gentry构造出了第一个真正的全同态加密体制

半同态加密 Partially Homomorphic Encryption (PHE)	浅同态加密 Somewhat Homomorphic Encryption (SWHE)	全同态加密 Fully Homomorphic Encryption (FHE)
仅支持加法同态（或乘法同态）的加密体制	同时满足加同态和乘同态性质，只能进行有限次的加和乘运算	同时满足加同态和乘同态性质，可以进行任意多次加和乘运算

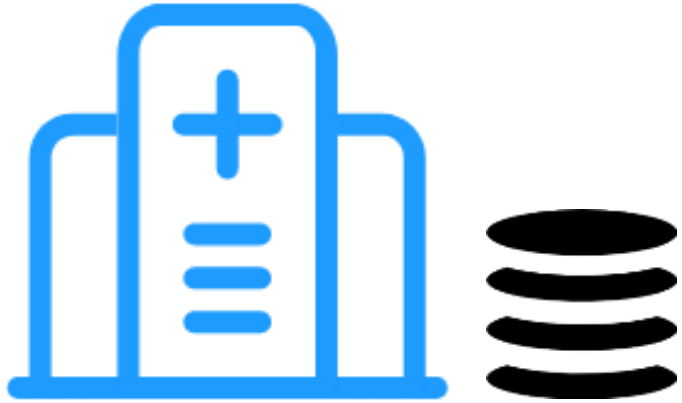
同态加密的发展

- 直到2009年，当时在斯坦福大学计算机科学系就读的博士生Craig Gentry才构造出了第一个真正的全同态加密体制
- 随后很多密码学家在全同态加密体制的研究方面取得进展，使得全同态加密继续向实用化靠近，但因当前算法复杂度问题，离实用仍有距离

实现了在不解密的条件下对加密数据进行任何可以在明文上进行的运算，对加密信息可以进行深入和无限的分析而不会影响其保密性

同态加密的应用

医疗机构的数据分析场景



如何在不泄露隐私的情况下
让云服务商存储和计算数据

数据处理能力较弱

云服务商提供计算服务

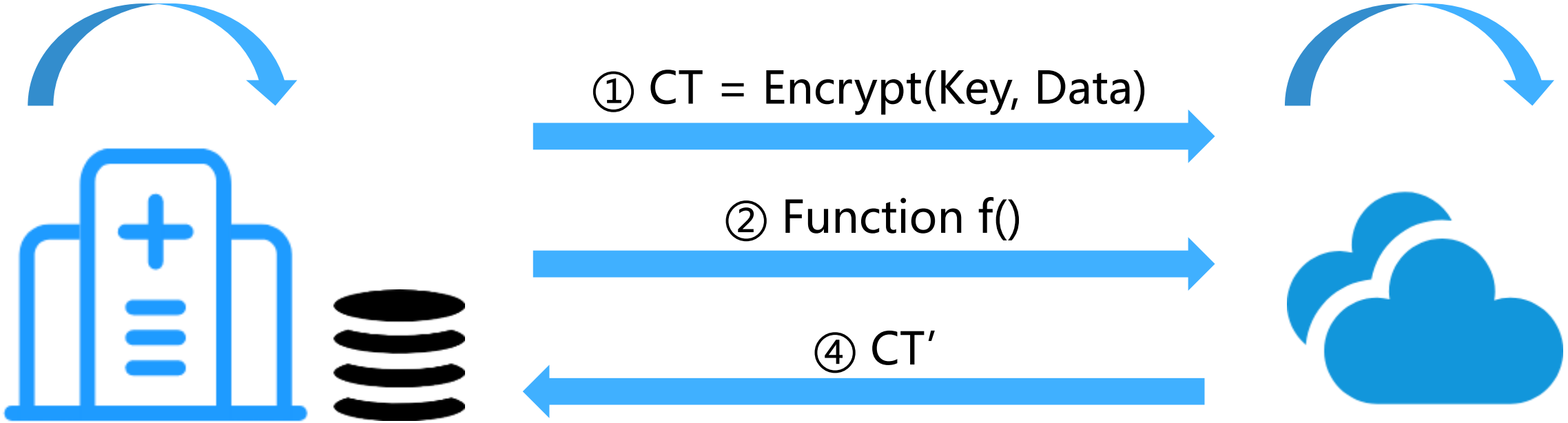
常见的方法是由用户对数据进行加密，把加密后的密文信息存储在服务端，但是传统的加密方案难以满足用户需要服务器提供数据搜索、分析、处理等功能时的需求，全同态加密可为这些功能的实现提供支持

同态加密的应用

医疗机构的数据分析场景

⑤ f(Data) = Decrypt(Key, CT′)

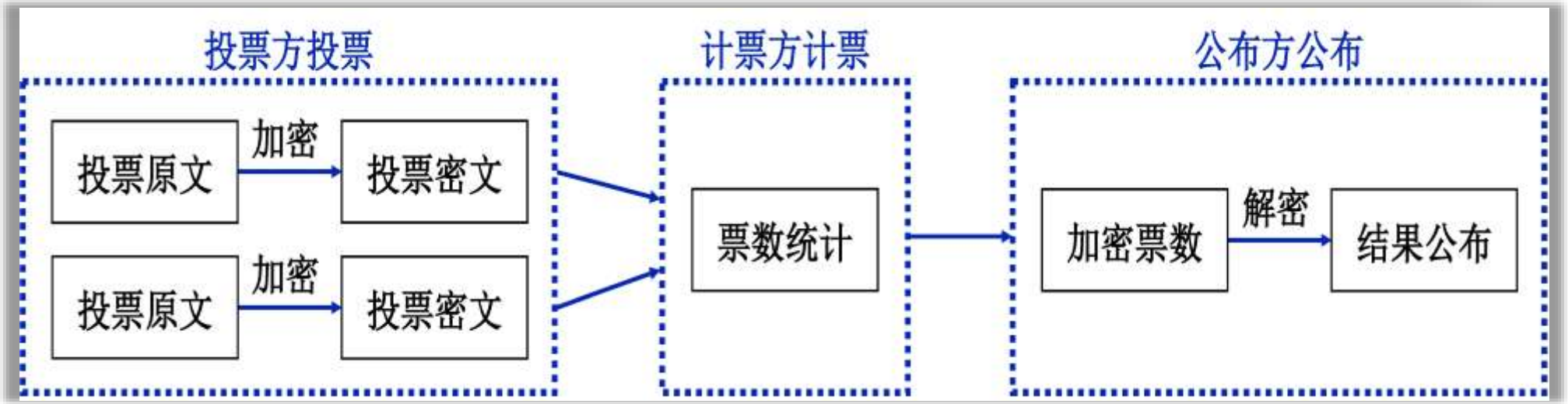
③ CT′ = Evaluate(f, CT)
= Encrypt(Key, f(Data))



同态加密的应用

- 与传统的投票方式相比，**电子投票**计票快捷准确，节省人力和开支，投票时具有便利性，而设计安全的电子选举系统是全同态加密的一个典型应用
- 基于同态加密设计的电子选举，统计方可以在不知道投票者投票内容的前提下，对投票结果进行统计，既保证了投票者的隐私安全，又能够保证投票结果的公证

电子投票的简化流程



同态加密的优势与挑战

优势

- 对多个密文进行计算后再解密，降低了计算代价
- 可实现无密钥方对密文的计算，密文计算无须经过密钥方，降低了通信代价
- 可实现让解密方只能获知最后的结果，而无法获得每一个密文的消息，可以保证信息的安全性（匿名电子投票）

挑战

- 只能实现单比特加密，如何高效地实现全同态加密有待进一步研究
- 大多基于未论证的困难问题，寻找可论证的困难问题依然是个难题
- 需要额外的消除噪音算法，如何设计一个具有自然同态性的全同态加密方案依然是一个问题

虽然同态加密正在逐步向实用化靠近，但是其安全性和实用性方面的研究还有很长的路要走

半同态加密

在一个加密方案中，用a、b表示明文，Enc表示加密算法，Dec表示解密算法，$\oplus$表示在明文域上的运算，$\otimes$表示在密文域上的运算，如果该加密方案中的加密算法和解密算法满足

$$Dec\big(Enc(a) \otimes Enc(b)\big) = a \oplus b$$

当$\oplus$表示乘法时，称该加密为乘法同态加密

当$\oplus$表示加法时，称该加密为加法同态加密

典型的乘法同态加密算法

- 1977年提出的RSA公钥加密算法
- 1985年提出的ElGamal公钥加密算法

典型的加法同态加密算法

- 1999年提出的Paillier公钥加密算法，是最常用且最具实用性的加法同态加密算法

ElGamal乘法同态加密

- 在1985年，Taher ElGamal基于有限域上的离散对数困难假设设计了ElGamal加密算法，该加密算法具有乘法同态性

离散对数问题

给定素数p，Z_p^*的一个生成元α，以及元素$\beta \in Z_p^*$，计算整数x，其中$0 < x \leq p-2$，满足$\beta \equiv \alpha^x (mod\ p)$成立

记群G上某ElGamal加密系统的公钥为$pk = (G, q, g, h)$，其中$h = g^x$，x为秘密私钥，对消息m实施ElGamal加密后的密文可表示为$E(m) = (g^r, m \cdot h^r)$，其中$r \leftarrow_R Z_q$，对于任意的明文消息$m_1$和$m_2$，ElGamal加密系统满足如下乘法同态性质：

$$E(m_1) \cdot E(m_2) = (g^{r_1}, m_1 \cdot h^{r_1})(g^{r_2}, m_2 \cdot h^{r_2}) = (g^{r_1+r_2}, (m_1 \cdot m_2)h^{r_1+r_2}) = E(m_1 \cdot m_2)$$

同理，ElGamal体制也支持任意次的乘法同态操作

Paillier加法同态加密

- 1999年由Pascal Paillier提出的，其安全性基于判定合数剩余类的问题
- 是第一种且应用最为广泛具有加法同态性的加密算法
- 已广泛应用在加密信号处理或第三方数据处理领域

判定合数剩余问题

令$N = pq$，其中，p和q为安全素数，任给定$z \in \mathrm{Z}_{N^2}^*$，判定z为N次剩余还是非N次剩余

N次剩余定义：给定$N = pq$，其中，p和q为安全素数，给定$z \in \mathrm{Z}_{N^2}^*$，若存在某个$y \in \mathrm{Z}_{N^2}^*$，使得$\mathrm{z} = y^N (mod N^2)$成立，则称$z$为（模$N^2$的）$N$次剩余；否则，则称$z$为（模$N^2$的）非$N$次剩余

Paillier加法同态加密

密钥生成

随机的选取两个大素数p和q，且满足$\gcd(pq,(p-1)(q-1))=1$

计算$N=pq$和$\lambda=lcm(p-1,q-1)$

选取随机数$g\in Z_{N^2}^*$（$Z_{N^2}^*$为小于N^2且与N^2互素的正整数集合），且能保证$\mu=\left(L(g^\lambda \bmod N^2)\right)^{-1} \bmod N$存在，其中$L(x)=\frac{x-1}{N}$

此时公钥为(N,g)，私钥为(λ,μ)

（lcm：求最小公倍数，gcd：求最大公约数）

为简化计算，随机的选取两个小素数$p=3,q=5$

则$N=pq=15$，$\lambda=lcm(p-1,q-1)=4$

随机选取$g=16$，且能保证$\mu=\left(L(g^\lambda \bmod N^2)\right)^{-1} \bmod N=(L(16^4 \bmod 225))^{-1} \bmod 15$

$=(L(61))^{-1} \bmod 15=4^{-1} \bmod 15=4$ 存在

最终得到公钥$(N,g)=(15,16)$，私钥$(\lambda,\mu)=(4,4)$

Paillier加法同态加密

公钥 $(N, g) = (15, 16)$ ，私钥 $(\lambda, \mu) = (4, 4)$

加密

已知明文$m \in Z_N$（Z_N为$\{0, \cdots, N-1\}$），选择随机数$r \in Z_N^*$，密文$c = g^m r^N \ mod \ N^2$

假设明文 $m = 7$，随机选择 $r = 2$

则密文$c = \ g^m r^N \ mod \ N^2 = 16^7 2^{15} \ mod \ 225 = 83$

解密

已知密文$c < N^2$，明文$m = L\left(c^\lambda \ mod \ N^2\right)\mu \ mod \ N$

已知密文$c = 83$

则明文$m = L\left(c^\lambda \ mod \ N^2\right)\mu \ mod \ N = L(83^4 \ mod \ 225) \cdot 4 \ mod \ 15 = L(196) \cdot 4 \ mod \ 15 = 7$

Paillier加法同态加密

记Paillier加密系统的公钥为$pk = (N, g)$，其中N为公开模，而g为公开基，对消息实m施Paillier加密后的密文可表示为$E(m) = g^m r^N mod N^2$，对于任意的明文消息m_1和m_2，Paillier加密系统满足如下**加法同态性质**：

$$E(m_1) \cdot E(m_2) = (g^{m_1} r_1^N)(g^{m_2} r_2^N) = g^{m_1+m_2}(r_1 r_2)^N = E(m_1 + m_2)$$

同理，Paillier也支持任意次加法同态操作，即对于任意消息$m_1, m_2, \cdots, m_n$，如下等式成立

$$E(m_1) \cdot E(m_2) \cdots E(m_n) = E(m_1 + m_2 + \cdots + m_n)$$

Paillier加法同态加密

验证加法同态性质

公钥 $(N, g) = (15, 16)$ ，私钥 $(\lambda, \mu) = (4, 4)$

假设明文$m_1 = 7$，随机选择$r_1 = 2$

密文$c_1 = g^{m_1} r_1^{N} \bmod N^2 = 16^7 2^{15} \bmod 225 = 83$

假设明文$m_2 = 2$，随机选择$r_2 = 7$

密文$c_2 = g^{m_2} r_2^{N} \bmod N^2 = 16^2 7^{15} \bmod 225 = 58$

两密文相乘$c = c_1\ c_2 = 4814$

解密得到明文

$$m = L\left(c^{\lambda} \bmod N^2\right)\mu \bmod N = L(4814^4 \bmod 225) \cdot 4 \bmod 15$$
$$= L\left((4814 \bmod 225)^4 \bmod 225\right) \cdot 4 \bmod 15 = L(89^4 \bmod 225) \cdot 4 \bmod 15$$
$$= L(91) \cdot 4 \bmod 15 = 9 = m_1 + m_2$$

全同态加密

全同态加密指同时满足加同态和乘同态性质，可以进行任意多次加和乘运算的加密函数，用数学公式来表达，即满足

$$Dec\left(f\left(Enc(m_1), Enc(m_2), \cdots, Enc(m_k)\right)\right) = f(m_1, m_2, \cdots, m_k)$$

或者写为$f\left(Enc(m_1), Enc(m_2), \cdots, Enc(m_k)\right) = En(f(m_1, m_2, \cdots, m_k))$

如果f是任意函数，称为全同态加密

- 鉴于全同态加密的强大功能，一经提出便成为密码界的公开问题，被誉为“密码学圣杯”
- 直到2009年，Gentry才基于理想格构造出了首个全同态加密方案，虽然在实际应用中效率不高，但这一里程碑事件激起了全同态加密研究的热潮

全同态加密方案

全同态加密	代表方案	提出时间	说明	分类
第一代	Gentry	2009	• 用理想格构造可保持对明文进行较低次数的多项式运算时的同态性的方案 • 给出一种将该方案修改为自举型方案的方法 • 证明了任意一个自举型方案都可以转换为全同态加密方案	无限层全同态加密方案： • 理论上可以进行无限深度的同态操作 • 同态操作的计算开销、密钥规模和密文尺寸都比较大
第二代	BGV	2012	• 突破了Gentry 的设计框架，在效率方面实现了很大的提升 • 在使用密钥交换技术时需要增加大量用于密钥交换的矩阵，从而导致公钥长度的增长	层次型全同态加密方案： • 需要预先给定所需同态计算的深度来执行该深度的多项式同态操作 • 可以满足绝大多数应用的需要
第三代	GSW	2013	• 被认为是目前最为理想的方案，不再需要密钥交换与模转换技术	

全同态加密与半同态加密

与半同态加密相比，全同态加密

- 加密算法功能更强大
- 具有较高的计算复杂度，加密算法设计更复杂
- 整体性能远不及半同态加密算法

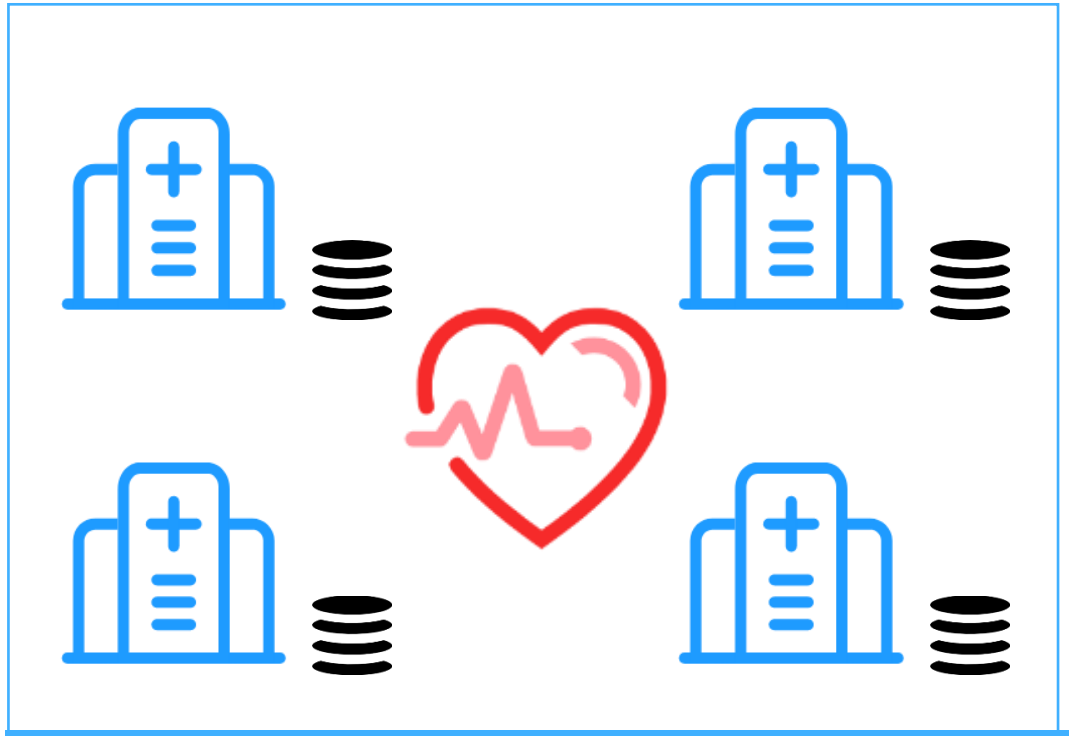
相关研究报告显示，在一次使用全同态加密开源库为敏感医疗数据构建密文线性回归模型的尝试中，1M的明文数据编码后可能膨胀至约10G密文数据

第5节 安全多方计算

- 安全多方计算基础
- 百万富翁协议

如何合作使用数据

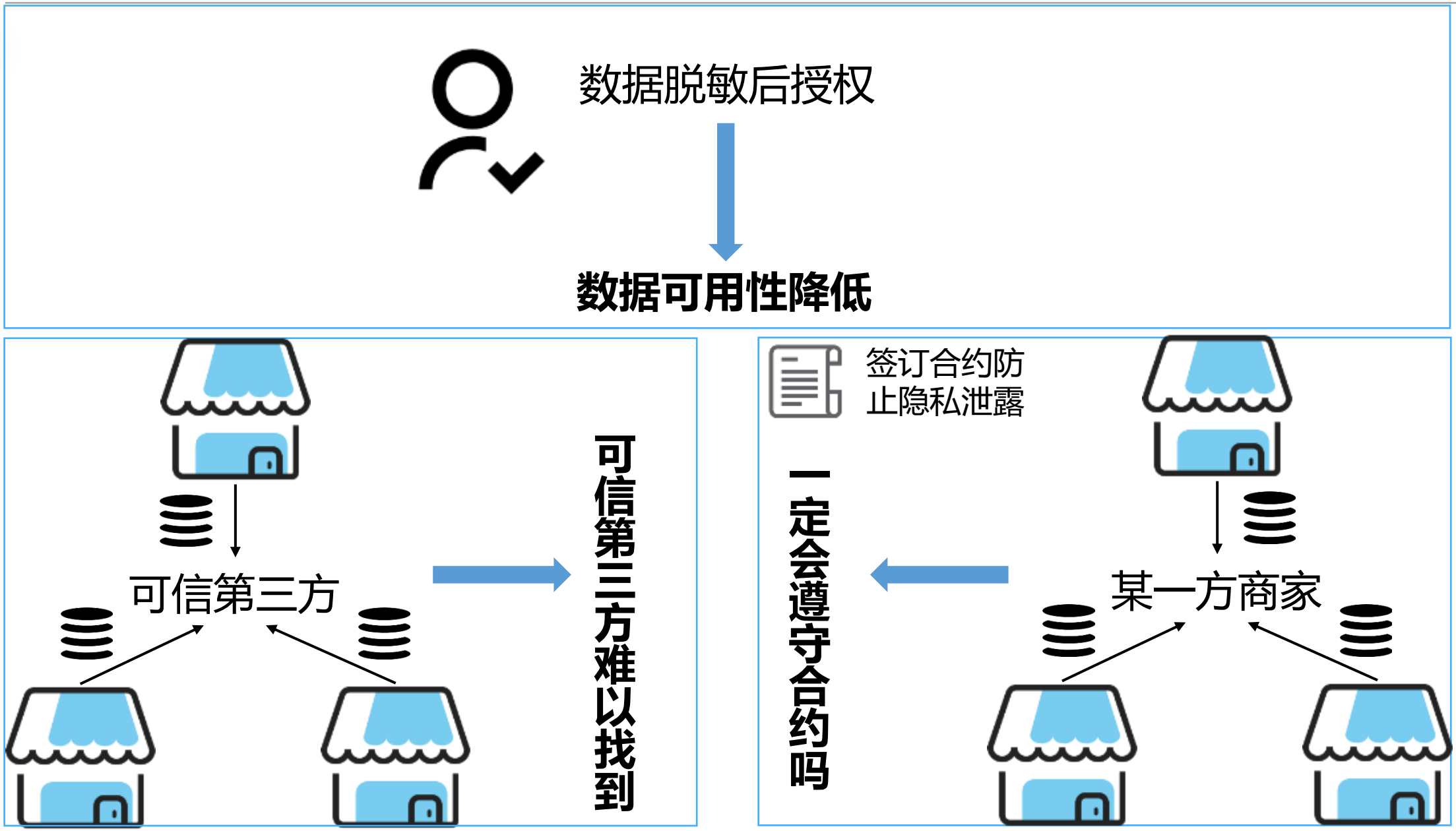
- 多家医院想要合作使用医疗数据进行科学研究、分析预测病人患病情况等，但为了保护患者隐私，不能直接共享数据



- 多个商家想要合作促销，统计共同的用户画像，但是又不想让对方知道自己掌握的信息

相互不信任的多方参与者如何在保护己方数据的前提下进行合作

如何合作使用数据

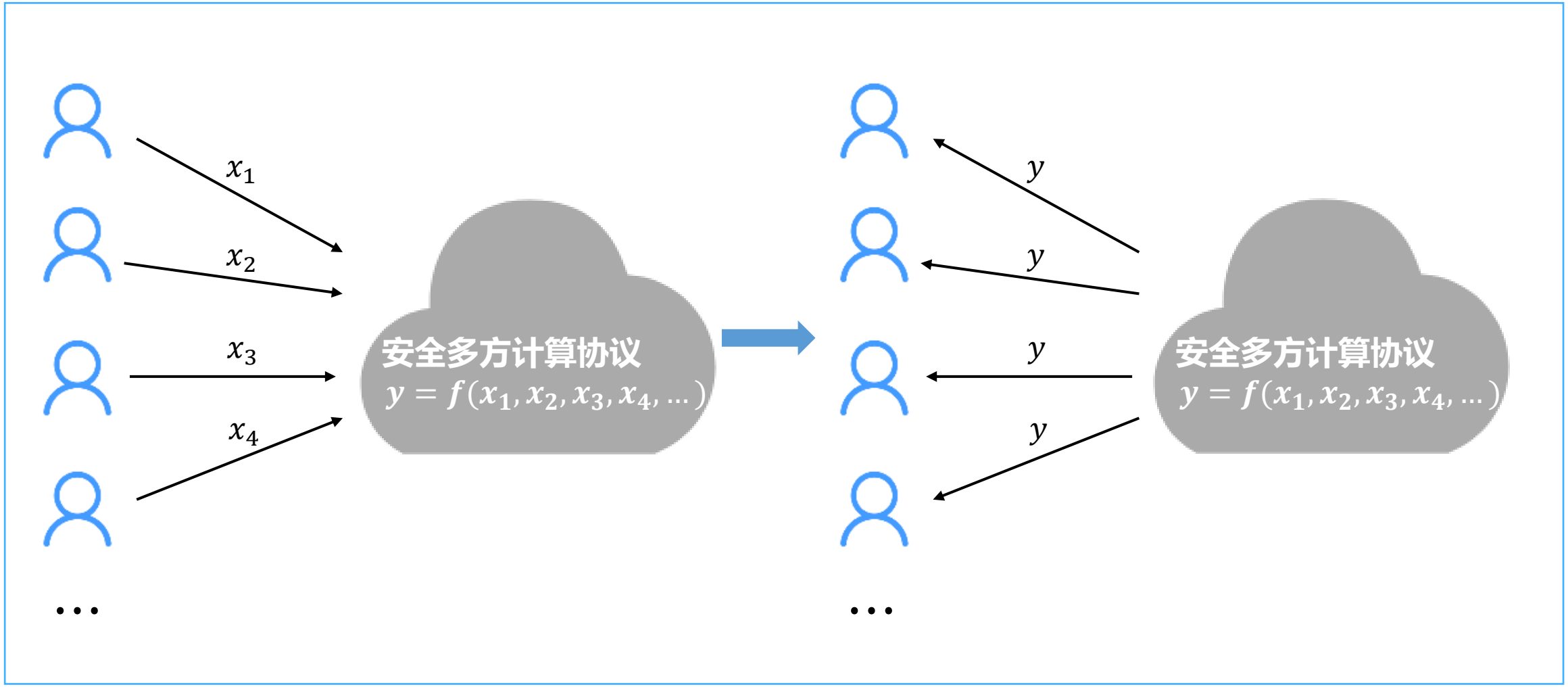

数据脱敏后授权
数据可用性降低
可信第三方
可信第三方难以找到
签订合约防止隐私泄露
某一方商家
一定会遵守合约吗

安全多方计算的提出

2000年图灵奖获得者

- 安全多方计算（**Multi-Party Computation, MPC**）起源于**姚期智教授**在1982年提出的百万富翁问题
- **解决了一组互不信任的参与方之间保护隐私的协同计算问题**
  - 当有两方或者多方参与者决定互相合作并且各自需要提供自己的隐私或者秘密数据时，任意一方都不愿意让其他一方知晓自己提供的信息

安全多方计算的提出


x_1
x_2
x_3
x_4
...
安全多方计算协议
$y = f(x_1, x_2, x_3, x_4, \ldots)$
y
y
y
y
...
安全多方计算协议
$y = f(x_1, x_2, x_3, x_4, \ldots)$

安全多方计算形式化描述

假定有m个参与方$P_1, P_2, \cdots, P_m$,他们拥有各自的数据集$d_1, d_2, \cdots, d_m$,在无可信第三方的情况下如何安全地计算一个约定函数$y = (d_1, d_2, \cdots, d_m)$，同时要求每个参与方除了计算结果外不能得到其他参与方的任何输入信息

安全多方计算的特征

输入独立性	需保证各方能独立输入数据，计算时不泄露本地数据
计算正确性	需保证计算结束后各方能够得到正确的计算结果
去中心化性	各参与方地位平等，提供了去中心化的计算模式

安全多方计算的威胁模型

现实世界中不存在

半诚实模型
在诚实模型基础上保留所有收集到的信息，推断其他参与者的秘密信息

多数情况符合

诚实模型
参与者不提供虚假数据，不会泄露、窃听数据，不会终止协议，完全按照协议执行

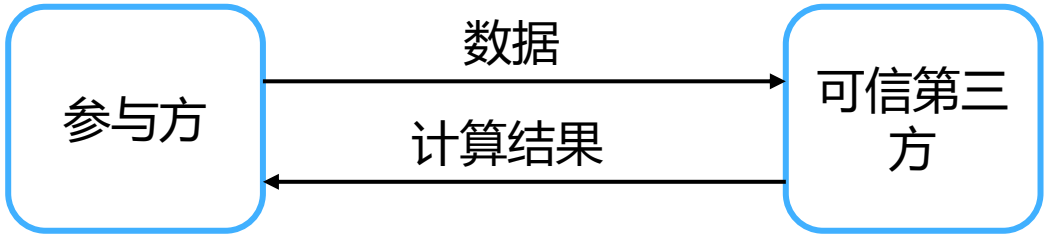
恶意模型
无视协议要求，可能提供虚假数据、泄露数据、窃听甚至终止协议

三种安全模型(根据参与方的可信程度)

计算模型

安全多方计算的计算模型主要有基于“可信第三方”的计算模型、交互计算模型和外包计算模型

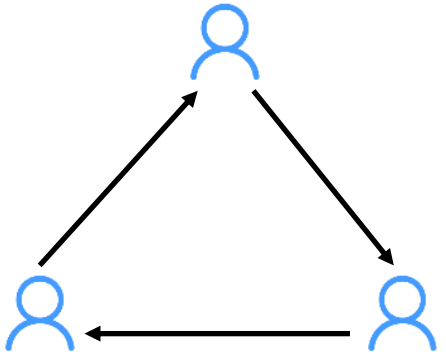
基于“可信第三方”的计算模型



- 参与者得到计算结果，可信第三方得到参与者的输入信息和计算结果，信息的保密性由可信第三方来保证

很难找到完全可信的第三方，不能满足实际中对安全性的需要，目前研究中很少使用

交互计算模型

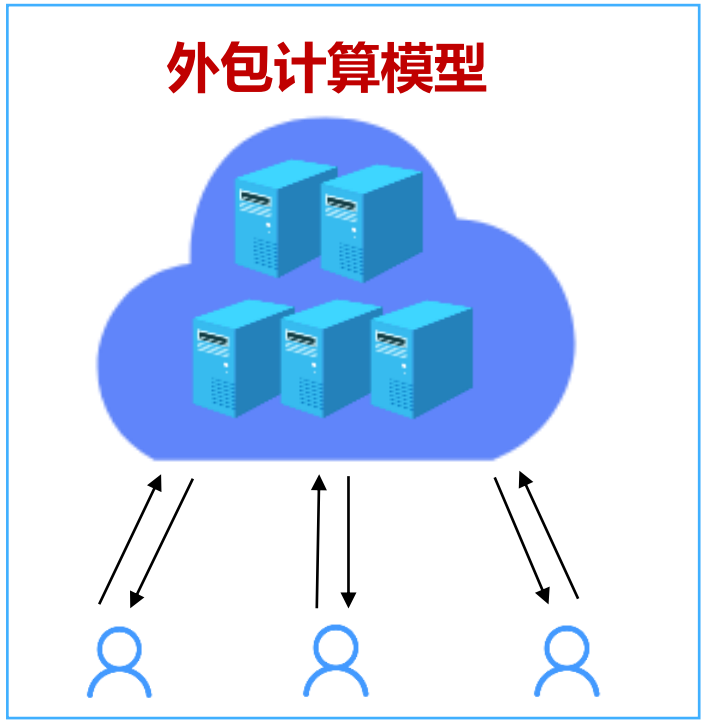


- 参与者约定协议通过交互计算共同完成函数运算
- 按照协议步骤执行计算，按协议的要求将中间结果发送给其他参与者同时接收其他参与者计算的中间结果，信息的保密性由协议的安全性来保证

使用最为广泛

计算模型

- 外包计算模型是随着**云计算**的发展而发展起来的计算模型
- 各个参与者希望在使用云计算提供的价格低廉的计算资源的同时不想直接将信息委托给云计算服务提供商，也不想让他们得知计算结果
  - 参与者将信息进行处理后存储在外包服务器上，由外包处理器对所有参与者的秘密信息进行计算,计算完成后将结果发送给各个参与者
  - 外包服务器可能需要和参与者进行必要的信息交换，信息的保密性由协议的安全性来保证



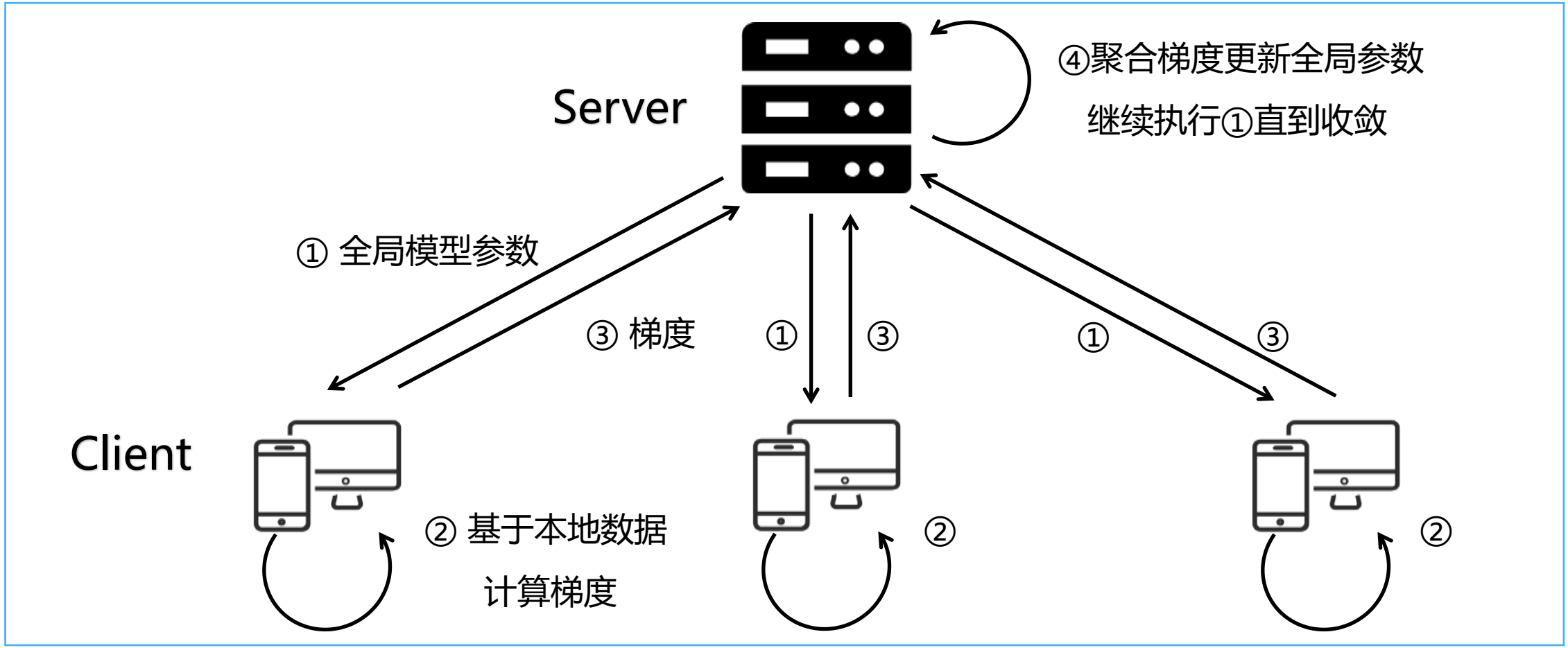
基本密码协议

- 安全多方计算基础密码协议包括茫然传输协议（Oblivious Transfer, OT）、混淆电路协议(Garbled Circuit, GC)、秘密共享协议（Secret Sharing）等，它们都是重要的密码学工具
- 安全多方计算是**多种密码学基础工具的综合应用**，在实现安全多方计算时也广泛地应用了同态加密技术

茫然传输协议	混淆电路协议	秘密共享协议
两方计算协议：接收方获得发送方发送的部分消息但不知道其他消息的内容，而发送方不知道哪些消息被接收	通用、高效的安全两方计算协议：两方能在互相不知晓对方数据的情况下计算某一能被逻辑电路表示的函数	将秘密以适当形式拆分并将拆分后的每一份交给不同的参与者管理，只有若干参与者相互协作才能恢复秘密

安全多方计算的应用

由于安全多方计算允许多个参与者通过使用各种加密技术,在不实际共享输入的地方对数据进行聚合计算，因此被重点应用在高效并行**分布式机器学习**中



安全多方计算的应用

• 门限签名	将私钥拆分为多个秘密分片，当不少于门限值的的秘密分片持有者共同协作时才能生成有效的签名
• 电子拍卖	需计算出所有参与者输入的最大值或最小值，安全多方计算理论的提出，使得网上拍卖成为现实
• 联合数据查询	不同数据库资源共享时，多个数据库可以看成多个用户联合起来进行数据查询，可使用安全多方计算保护各数据库的私有信息或知识版权

- 安全多方计算牵涉到密码学的各个分支，有着广阔的应用领域
- 其优势为比较安全和准确，但涉及的加密技术开销、通信开销也很大
- 目前的研究主要集中于降低计算开销、优化分布式计算协议

百万富翁问题

- 姚期智教授1982年提出的百万富翁问题是**第一个安全双方计算问题**
- 百万富翁Alice和Bob想相互比较一下谁更富有，但是他们都不想让对方知道自己拥有多少财富，如何不借助第三方比较两个人的财富多少?

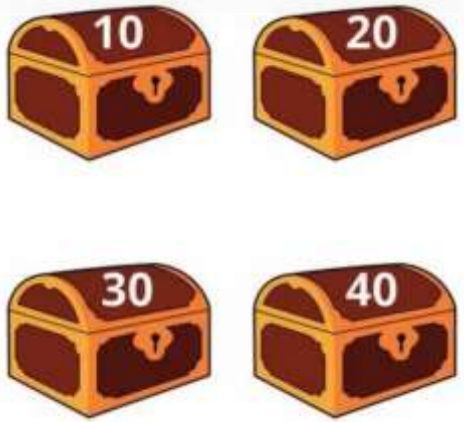
在不知道对方财富的情况下比较谁更富有

百万富翁问题 – 直观解决方法

存在可信第三方

- 当不存在可信的第三方，天平谁来提供?
- 提供天平者是可以知道一切的

双方的行为可信



- 如何确保双方根据协定来执行协议?

双方的行为互不可信

- 如何通过应用密码学技术实现百万富翁协议?

百万富翁协议

输入：Alice和Bob的财富值 i，j，Alice拥有公私钥

输出：$i \geq j$ 或 $i < j$

- 设Alice拥有的财富为i,Bob拥有的财富为j，单位均为百万，其中$1 \ll \mathrm{i}, \mathrm{j} \ll 10$
- 令M为N 个bit表示的非负整数的集合，Q_N是从M映射到M的所有一一映射的集合
- E_a是Alice的公钥，通过从Q_N中随机选择一个元素生成，D_a为私钥

百万富翁协议

Alice

2. Bob将$k-j+1$发送给Alice

Bob

1. Bob选择一个$N\ bit$的随机整数x，并私下计算$k=E_a(x)$

3. Alice计算$Y_u = D_a(k-j+u)$的值，其中$u=1,2,\cdots,10$

Bob知道解密后的序列中第j个数为随机数x，因为当$u=j$时，$D_a(k-j+u)=D_a(k)=\mathrm{x}$，但是由于没有私钥，因此不知道其他解密值

百万富翁协议

Alice

5. Alice对序列Z_u进行处理，将p和序列$z_1, z_2, \cdots, z_i, z_{i+1}+1, \cdots, z_{10}+1$发送给Bob

Bob

4. Alice生成一个$\frac{N}{2}$ bit的随机

$Y_u(mod\ p)$，该随机素

第j个数字,如果该数字等于

$< j$

- Alice从序列中的第$i+1$个数开始对数值做加一处理，当$i \geq j$时，第j个数值不变，当$i < j$时，第j个数值被修改，因此Bob可通过判断第j个数值是否被修改来比较i和j的大小
- 由于Bob没有私钥，不知道Y_u的解密值，因此无法知道除第j个数以外的其他数是否被修改，也就无法判断i的大小

百万富翁协议后续

- 姚期智教授在提出安全多方计算后，提出了基于混淆电路的通用解决方案，验证了安全多方计算的通用可行性，奠定了现代计算机密码学的理论基础
- 混淆电路和茫然传输成为了主流安全两方计算框架的核心
- 此后经过密码学者的进一步研究和创新，安全多方计算逐渐发展成为了现代密码学的一个重要分支

How to Generate and Exchange Secrets
(extended abstract)

Andrew Chi-Chih Yao*
Department of Computer Science
Princeton University
Princeton, New Jersey 08544

Abstract

In this paper we introduce a new tool for controlling the knowledge transfer process in cryptographic protocol design. It is applied to solve a general class of problems which include most of the two-party cryptographic problems in the literature. Specifically, we show how two parties A and B can interactively generate a random integer $N = p \cdot q$ such that its secret, i.e., the prime factors (p, q), is hidden from either party individually but is recoverable jointly if desired. This can be utilized to give a protocol for two parties with private values i and j to compute any polynomially computable functions $f(i,j)$ and $g(i,j)$ with minimal knowledge transfer and a strong fairness property. As a special case, A and B can exchange a pair of secrets x_A, x_B, e.g. the factorization of an integer and a Hamiltonian circuit in a graph, in such a way that x_A becomes computable by B when and only when x_B becomes computable by A. All these results are proved assuming only that the problem of factoring large integers is computationally intractable.

1. Introduction.

A protocol $\Pi = (M_A, M_B)$ is a pair of communicating probabilistic Turing machines such with a special "send-receive" tape. Given inputs of the form (n, i_A) and (n, i_B), the two machines will alternately send and receive message strings using the send-receive tapes; each machine will perform computation as a standard Turing machine after receiving a message string, including the computation of the next message to be sent. Eventually both machines halt within a number of steps bounded by some polynomial in n, leaving strings o_A and o_B on the output tapes. For a detailed description, see for example [GMR] [GHY]; in our case we allow each machine to have their own private input tape. For any run σ of the protocol, let $\Delta_A(\sigma)$ denote the history of the run from A's view, i.e. a sequence of the instantaneous descriptions of M_B; similarly $\Delta_B(\sigma)$ denote the history of this run from B's view.

* This research was supported in part by National Science Foundation under grants DCR-83-00990.

0272-5428/86/0000/0162$00.00 © 1986 IEEE

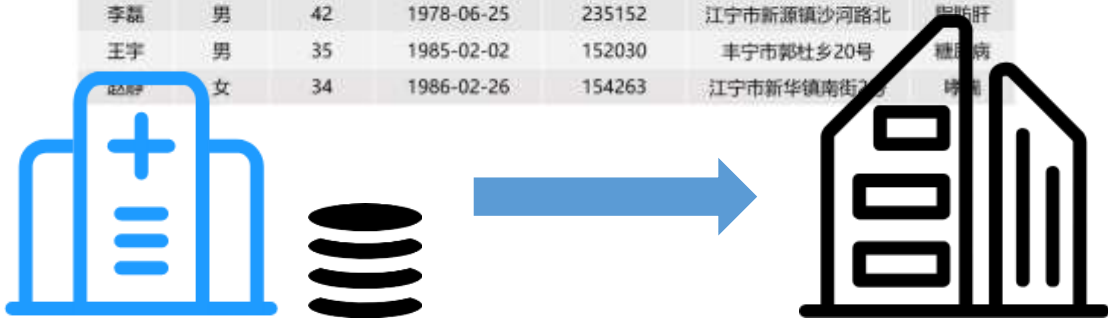
162

第4节 总结和展望

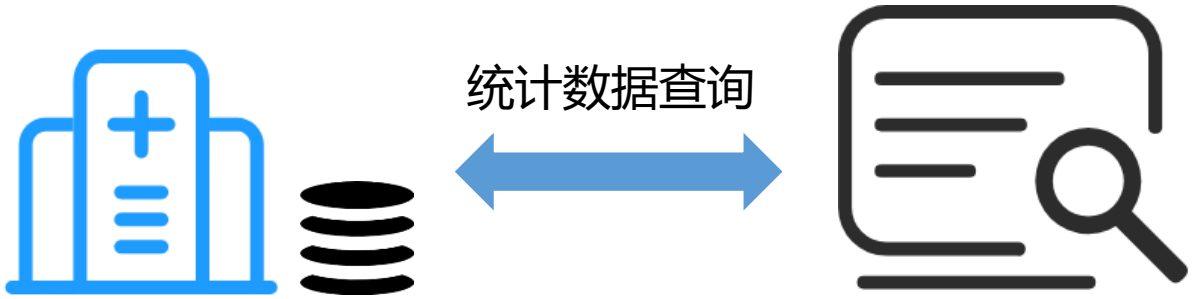
多种场景下的隐私泄露与保护问题

病患记录表

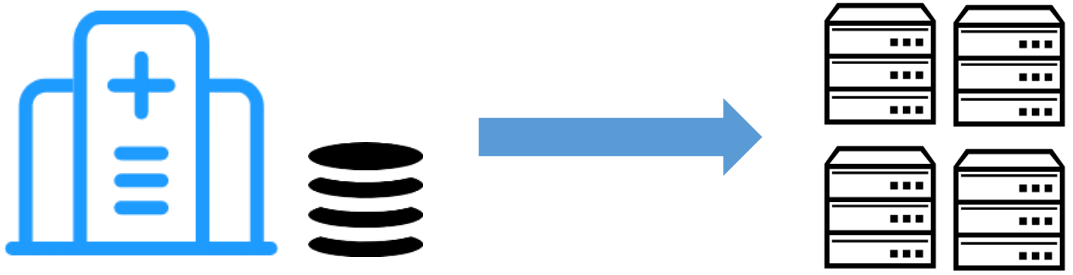
姓名	性别	年龄	出生日期	邮政编码	家庭住址	疾病
张艳	女	36	1984-12-03	235023	海河市江流镇北路1号	流感
李磊	男	42	1978-06-25	235152	江宁市新源镇沙河路北	脂肪肝
王宇	男	35	1985-02-02	152030	丰宁市郭杜乡20号	糖尿病
赵静	女	34	1986-02-26	154263	江宁市新华镇南街2号	哮喘



医院发布数据供其他机构研究，患者信息可能被泄露，需要对数据进行匿名化



医院数据库对外提供统计查询服务，患者信息可能被泄露，使用差分隐私保护个体隐私



医院委托其他机构对数据进行计算，患者信息可能被泄露，使用同态加密保护患者隐私

有多家医院想要合作进行科学研究，患者信息可能被泄露，使用安全多方计算帮助多家医院进行协同计算

总结

概述数据隐私保护技术，讲解不同分类下的隐私保护技术思想与基础知识

第一节
隐私保护技术初探

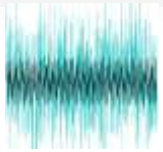
- 网络空间安全中的隐私
- 隐私泄露的危害
- 隐私保护技术介绍

多种数据隐私保护技术保护了网络空间中的数据隐私

第二节
匿名化

- 匿名化隐私保护模型
- 数据匿名化方法

通过隐藏用户身份和数据的对应关系来保护隐私



第三节
差分隐私

- 差分隐私基础
- 数值型差分隐私
- 非数值型差分隐私

通过在查询结果中添加噪声来保护隐私

第四节
同态加密

- 同态加密基础
- 半同态加密
- 全同态加密

通过加密数据来保护计算过程的安全

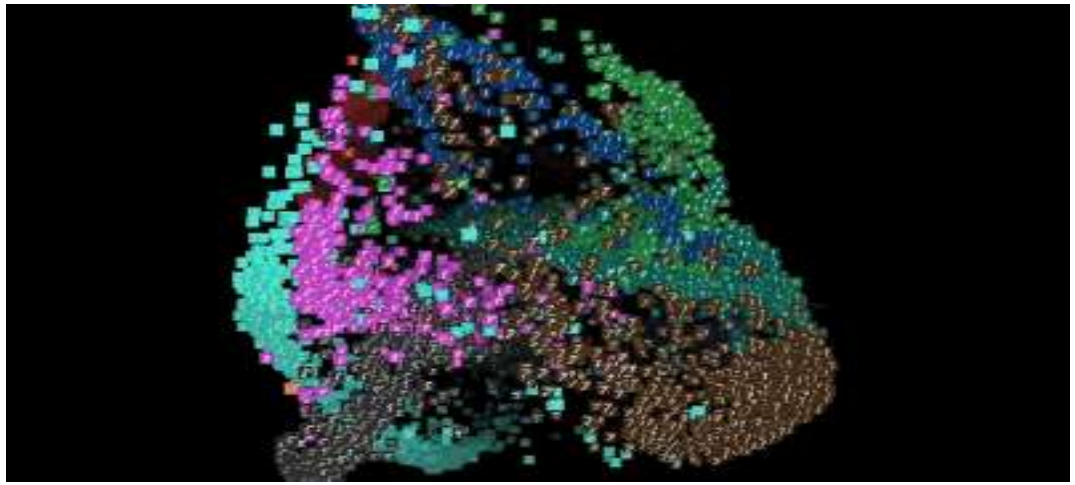
第五节
安全多方计算

- 安全多方计算基础
- 百万富翁协议

解决互不信任参与方之间进行协同计算的隐私保护框架

展望

研究如何更好的保护动态数据、高维数据中的隐私



- 数据隐私保护技术在静态数据的隐私保护方面比较成熟，而对于**如何保护动态数据中的隐私**还存在较多问题
- 实际中的高维数据越来越多，使用传统的隐私保护模型处理高维数据会导致信息损失过多，**如何保护高维数据的隐私**也是未来需要研究的问题