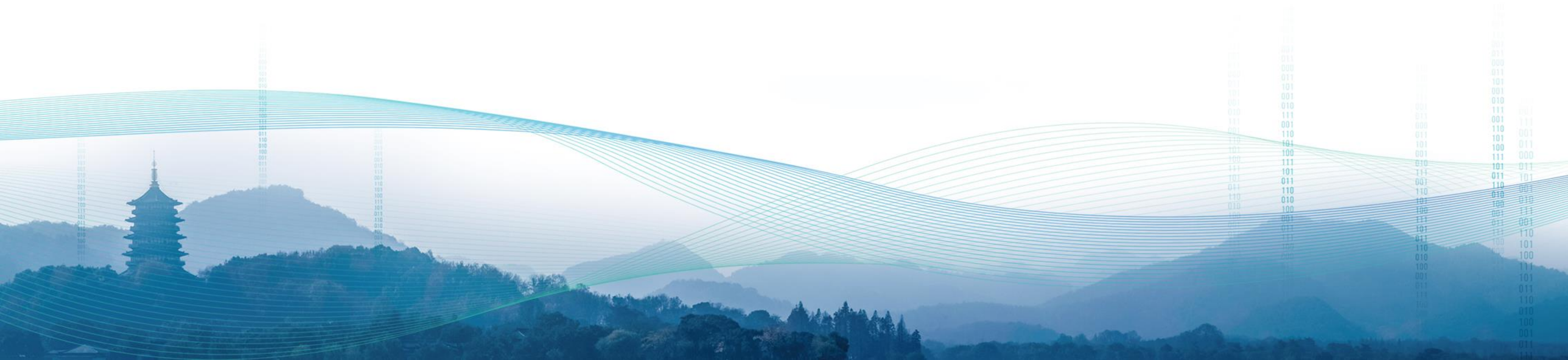
2019 西湖论剑·网络安全大会

WEST LAKE CYBERSECURITY CONFERENCE

大数据交易与处理中的数据脱敏技术研究

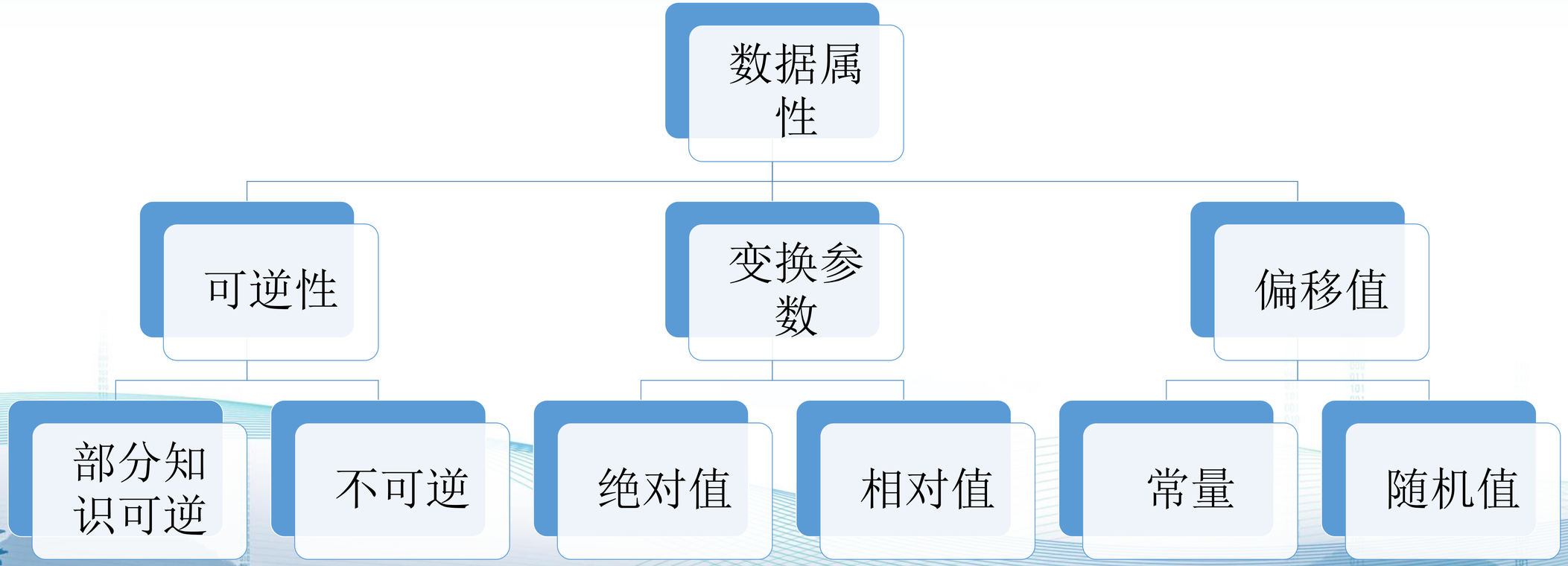
荆 继 武

CONTENTS
目 录



数据脱敏的指标

数据脱敏的有效性从可逆性体现，数据方法可通过变换参数和变换偏移值体现



数据脱敏的指标

数据脱敏的有

参数和变换偏移值体现

得知部分初始数据、或可逆的脱敏方法、或脱敏使用的伪随机数生成器及种子，可推演出原始数据

例：$y_i = x_i + constant$

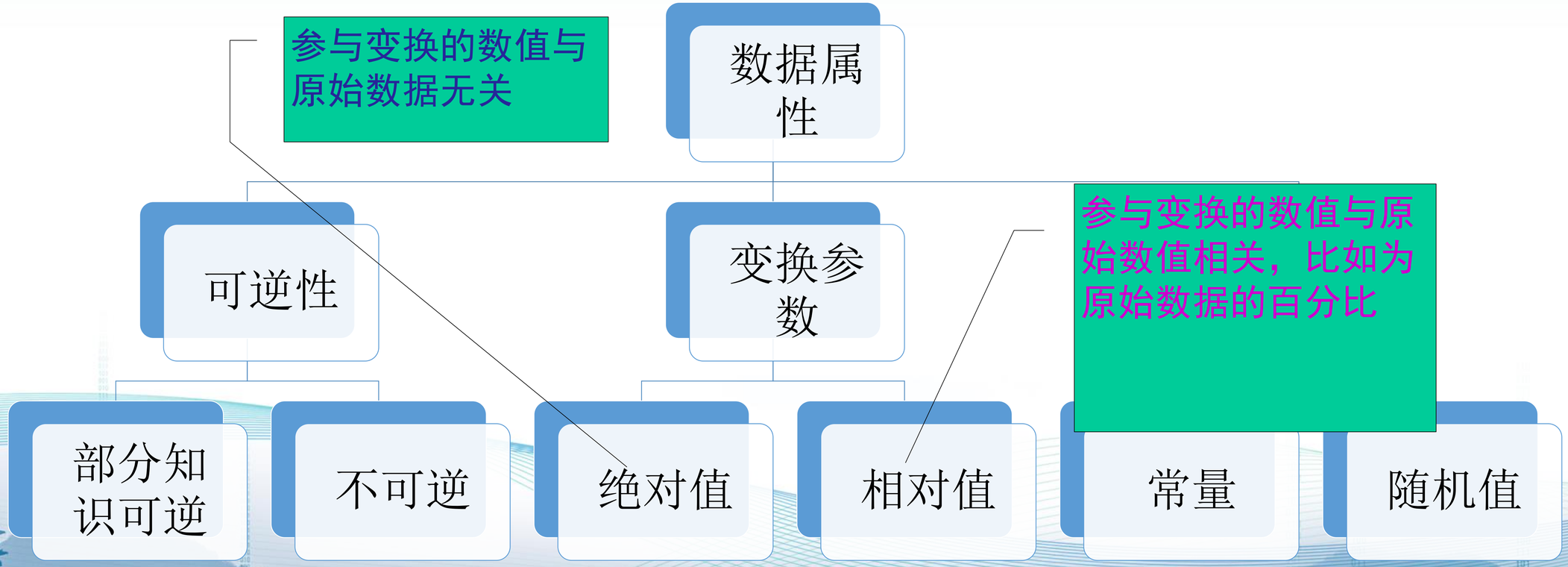
$y_i = f(x_i)$

$y_i = x_i + random\_number$

- 可逆性
  - 部分知识可逆
  - 不可逆
- 绝对值
- 相对值
- 偏移值
  - 常量
  - 随机值

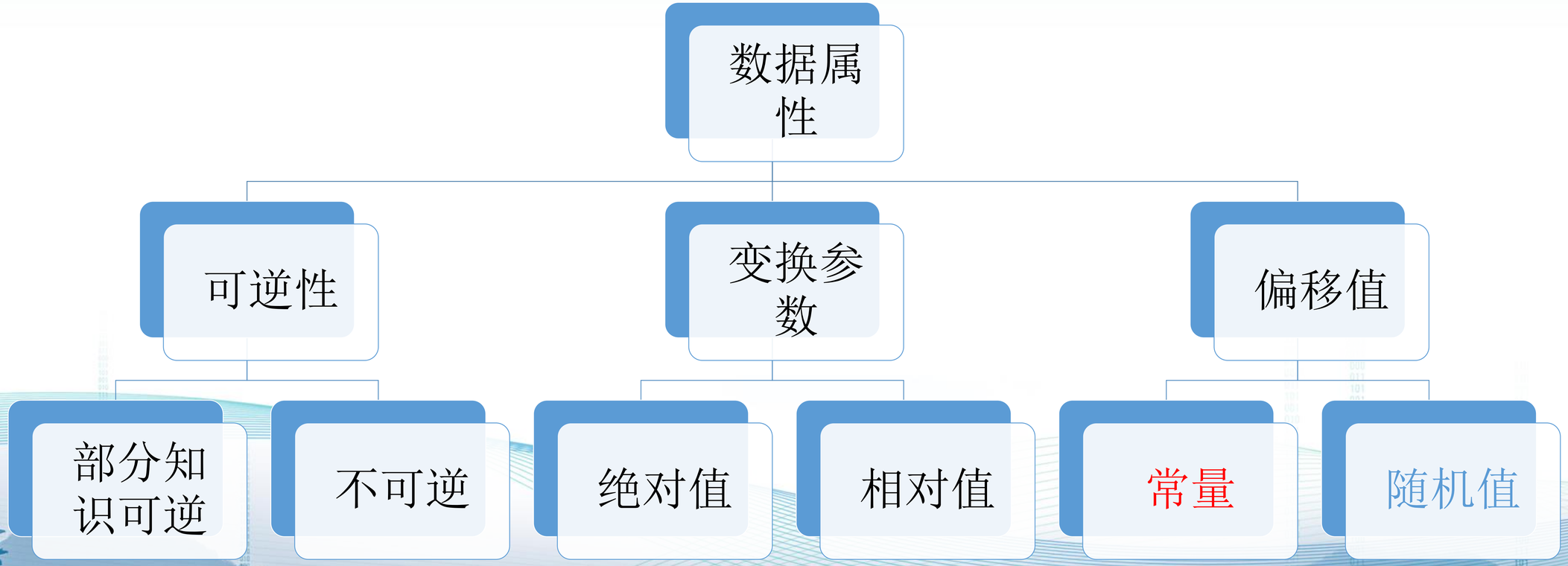
数据脱敏的指标

数据脱敏的有效性从可逆性体现，数据方法可通过变换参数和变换偏移值体现



数据脱敏的指标

数据脱敏的有效性从可逆性体现，数据方法可通过变换参数和变换偏移值体现



基于失真的数据脱敏方法

- 最终用户关注数据的聚合结果，不关注个体数据
  - 聚合结果：患某种疾病的人数
  - 个体数据：某个病人患该疾病

- 问题：提取聚合结果的时候可能披露个体数据
  - 患某种疾病的人数为N
  - 病人名字不为A，患某种疾病的病人的人数为M

- 基于失真的数据脱敏技术：在破坏个体隐私数据的基础上，不影响数据的聚合结果
  - 阻塞
  - 随机化

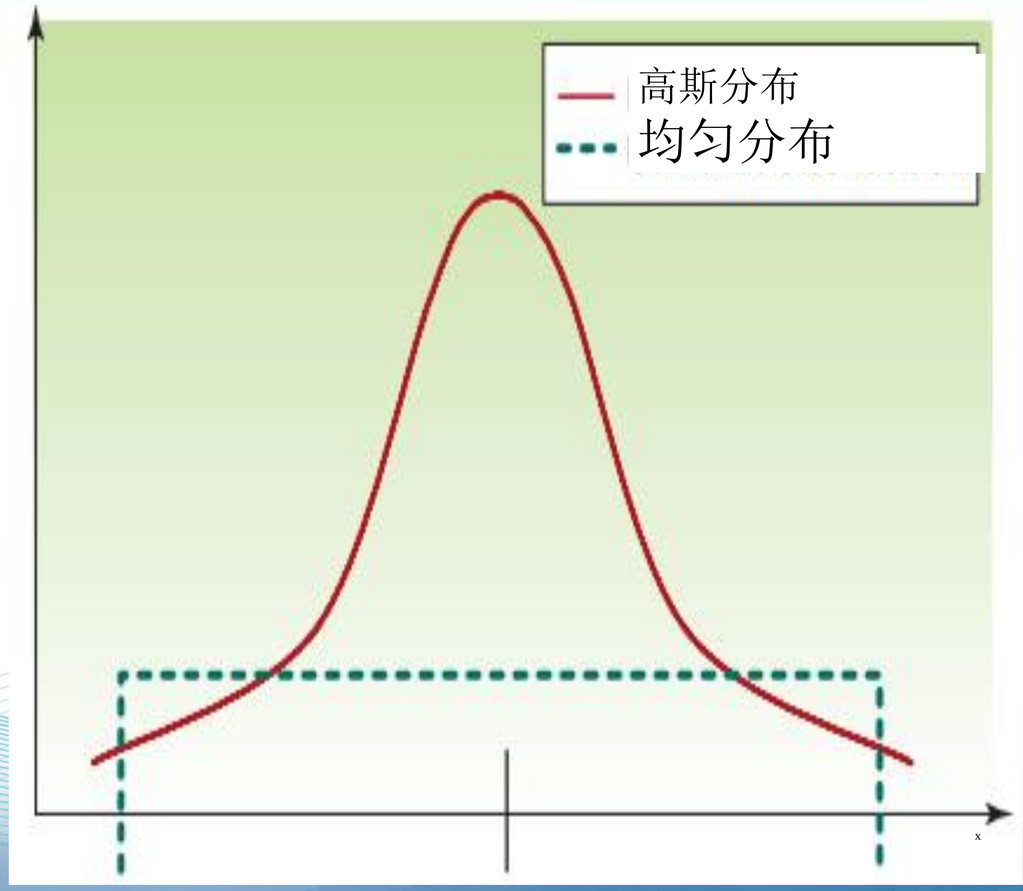
基于失真的数据脱敏方法：泛化

- **对原始数据不引入虚假噪声，仅泛化处理**
  - **典型方法1：离散化**
    - 属性值被离散化到各个区间
    - 区间大小不能等长
    - 使用区间作为属性来参与运算
    - 如：张三的年龄为25岁，使用区间[20,30]表征张三的年龄
  - **典型方法2：使用“?”替代数据中的某些属性**

- 同一区间内的值表征形式一致，脱敏后聚合准确率低
- 不同应用需要设计特定算法对处理后的数据进行处理

基于失真的数据脱敏方法：随机化

- 随机化
  - 实际数据：x_i
  - 使用 $x_i + r$，r是符合某个分布的随机值
    - 均匀分布
      - r 均匀分布于$[-\alpha, +\alpha]$，平均值为0
    - 高斯分布
      - r 符合高斯分布
      - 均值 $\mu(r)$ 为0
      - 标准方差为 σ



脱敏后的源数据分布重构

- 定义：
  - 原始数据值： x_1， x_2， …， x_n
  - 随机失真变量： y_1， y_2， …， y_n
  - 失真样本： x_1+y_1， x_2+y_2， …， x_n+y_n
  - F_Y： 随机失真变量y_i的累计分布函数CDF
  - F_X： 原始数据值x_i的累计分布函数CDF
- 重构问题：
  - 给定失真样本(x_1+y_1, …， x_n+y_n)， F_Y
  - 估算F_X

脱敏后的源数据分布重构算法

(1) f_X^0 := Uniform distribution
(2) $j := 0$ // Iteration number
repeat
(3) $f_X^{j+1}(a) := \frac{1}{n}\sum_{i=1}^{n} \frac{f_Y(w_i-a)\, f_X^j(a)}{\int_{-\infty}^{\infty} f_Y(w_i-z)\, f_X^j(z)dz}$
(4) $j := j+1$
until (stopping criterion met)

使用贝叶斯定律运算F_X:

1.初始化f(x, 0)：均匀分布

2.自j=0到终止条件

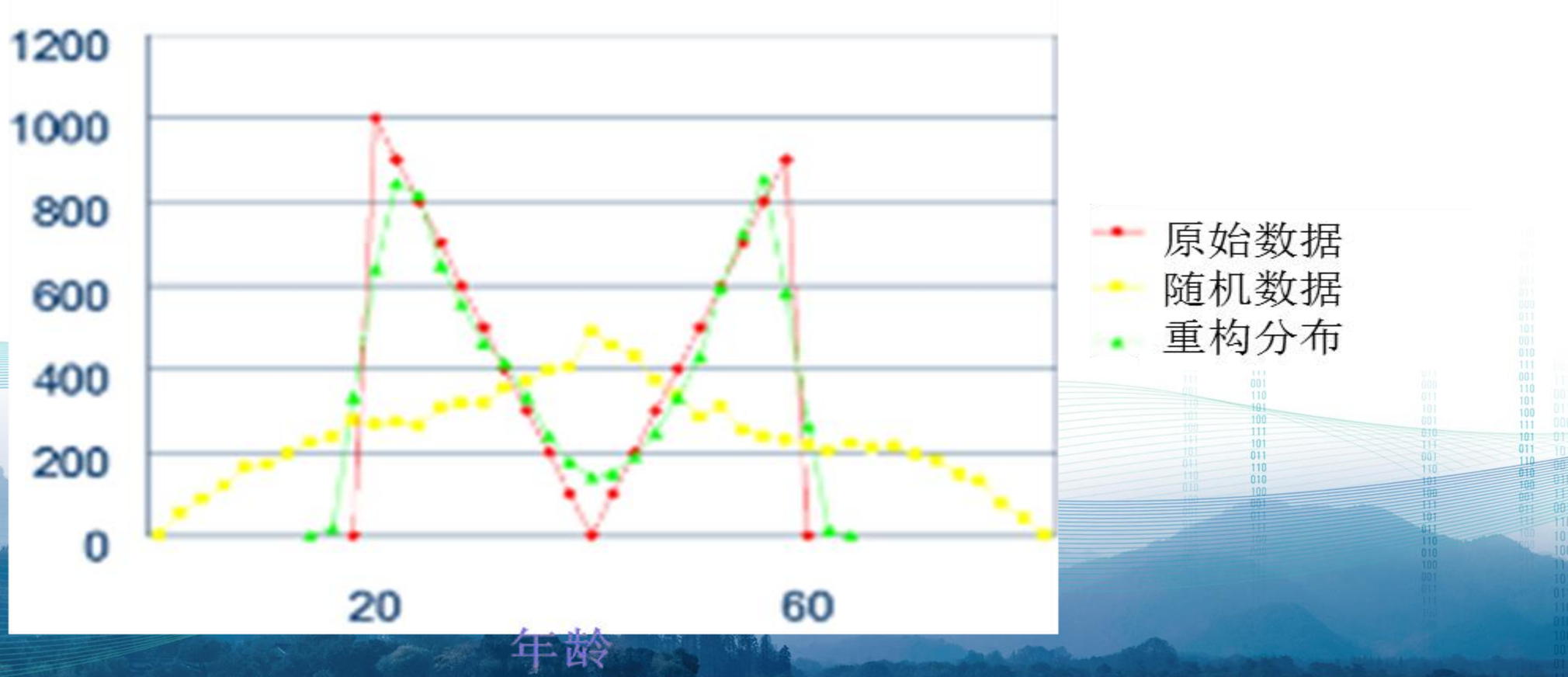
3.根据f(x, j) 和 F_Y 计算f(x, j+1)

4.满足条件终止，得到 F_X

终止条件：
1. 计算f(x, j).
2. 当f(x, j+1)与f(x, j)之间的差值非常小时

脱敏后的源数据分布重构实验

实验结果表明：重构后的数据分布与原始数据分布基本一致，即使随机数据样本分布与原始数据相差甚远



随 机 化 数 据 脱 敏 总 结

- 通过添加随机噪声扰乱失真敏感数据
  - 随机数必须随机！分布必须准确！
- 原始值未知，以保护数据敏感信息
- 数据脱敏后，能够准确获得聚合分类结果（支持决策树等）
- 有实验认为：在高置信度的情况下，高斯分布的随机噪声比均匀分布效果好

- 其他相关研究
  - 期望最大化（Expectation Maximization）算法

基 于 加 密 的 数 据 脱 敏

- 同态加密算法：
  - A way to delegate processing of your data, without giving away access to it.（Craig Gentry）
  - 他人可对加密数据进行处理，但处理过程中不会泄露原始数据

- 基于同态加密的数据脱敏技术：
  - 用户将数据进行同态加密后，提交给数据中心存储
  - 数据中心需要对数据进行分析处理时，可在不知道用户数据的前提下正确处理数据

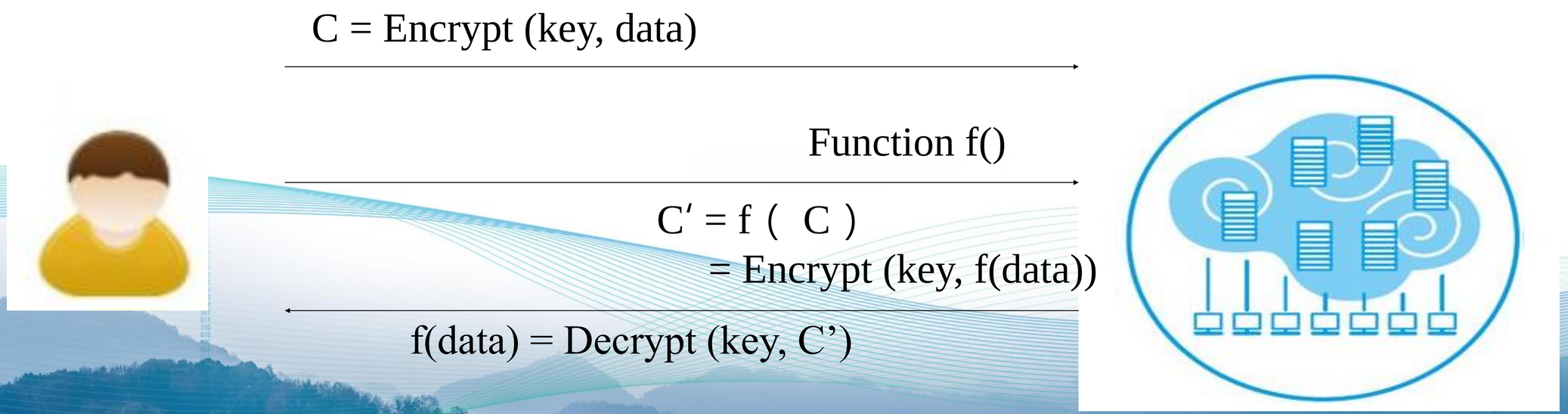
基 于 加 密 的 数 据 脱 敏

- 同态加密算法：
  - A way to delegate processing of your data, without giving away access to it.（Craig Gentry）
  - 他人可对加密数据进行处理，但处理过程中不会泄露原始数据

- 基于同态加密的数据脱敏技术：
  - 用户将数据进行同态加密后，提交给数据中心存储
  - 数据中心需要对数据进行分析处理时，可在不知道用户数据的前提下正确处理数据

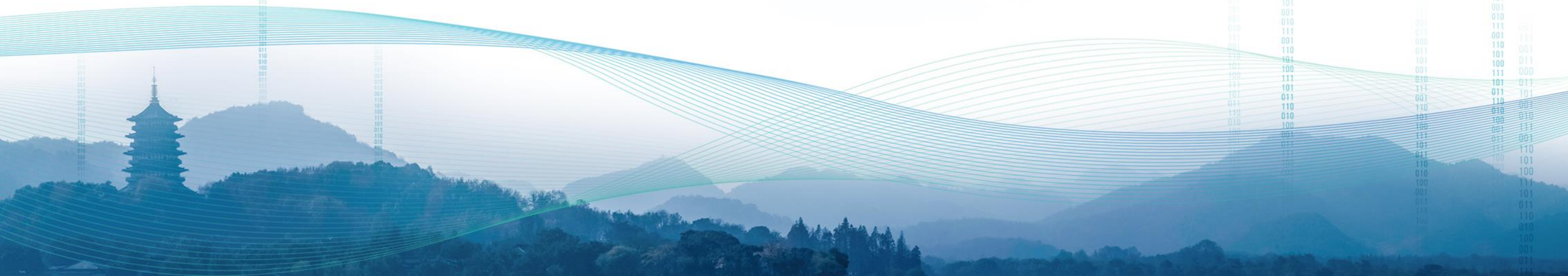
基于加密的数据脱敏：同态加密

- 密钥生成：key
- 加密函数：加密用户数据，生成密文
- 评估函数：在给定数据处理函数f下，对密文进行操作，使得结果相当于用户用密钥key对f(data)进行加密
- 解密函数：用于获取处理结果f(data)



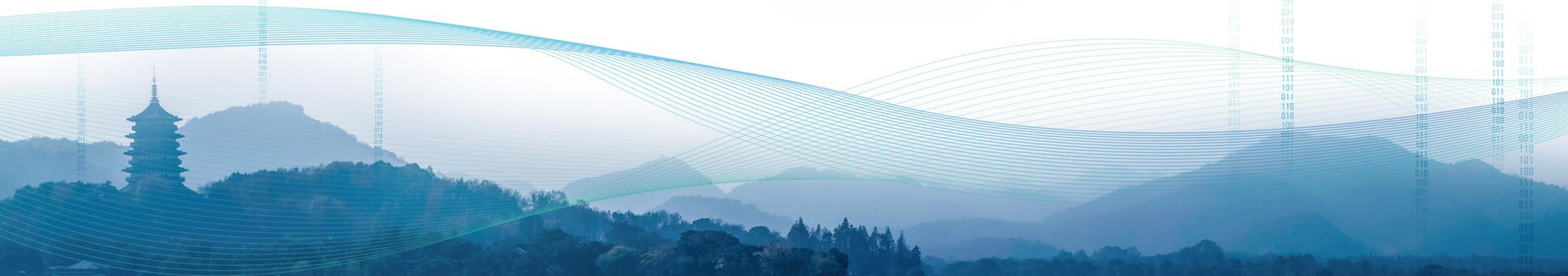
基于加密的数据脱敏：同态加密

- 全同态加密：
  - 支持任意给定的数据处理函数f，脱敏后的数据可满足任意数据处理需求
  - 开销大，难以满足实际应用
- 部分同态加密：
  - 支持特定的数据处理函数f，即脱敏后的数据只能满足特定的数据处理需求
  - 开销小，易实现，已可在实际应用中使用



基于加密的数据脱敏：同态加密

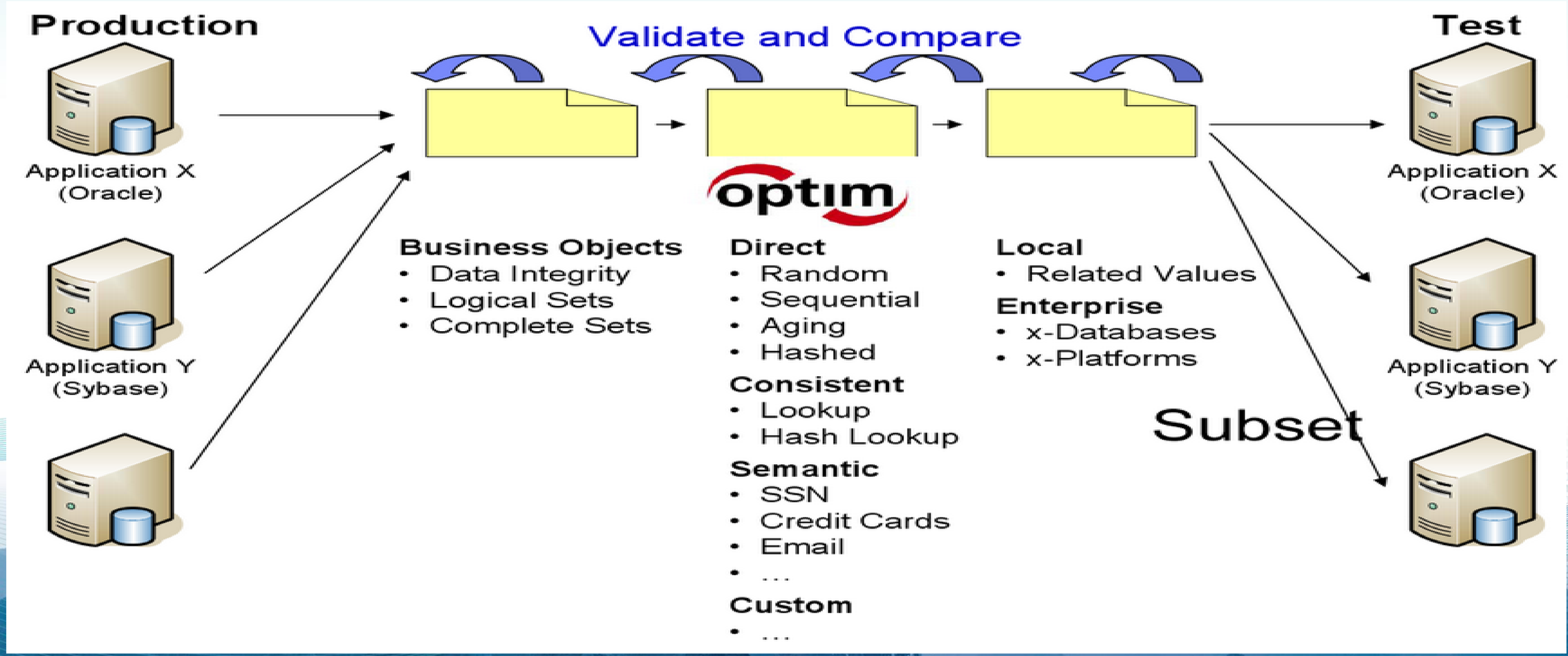
- 全同态加密：
  - 支持任意给定的数据处理函数f，脱敏后的数据可满足任意数据处理需求
  - 开销大，难以满足实际应用
- 部分同态加密：
  - 支持特定的数据处理函数f，即脱敏后的数据只能满足特定的数据处理需求
  - 开销小，易实现，已可在实际应用中使用



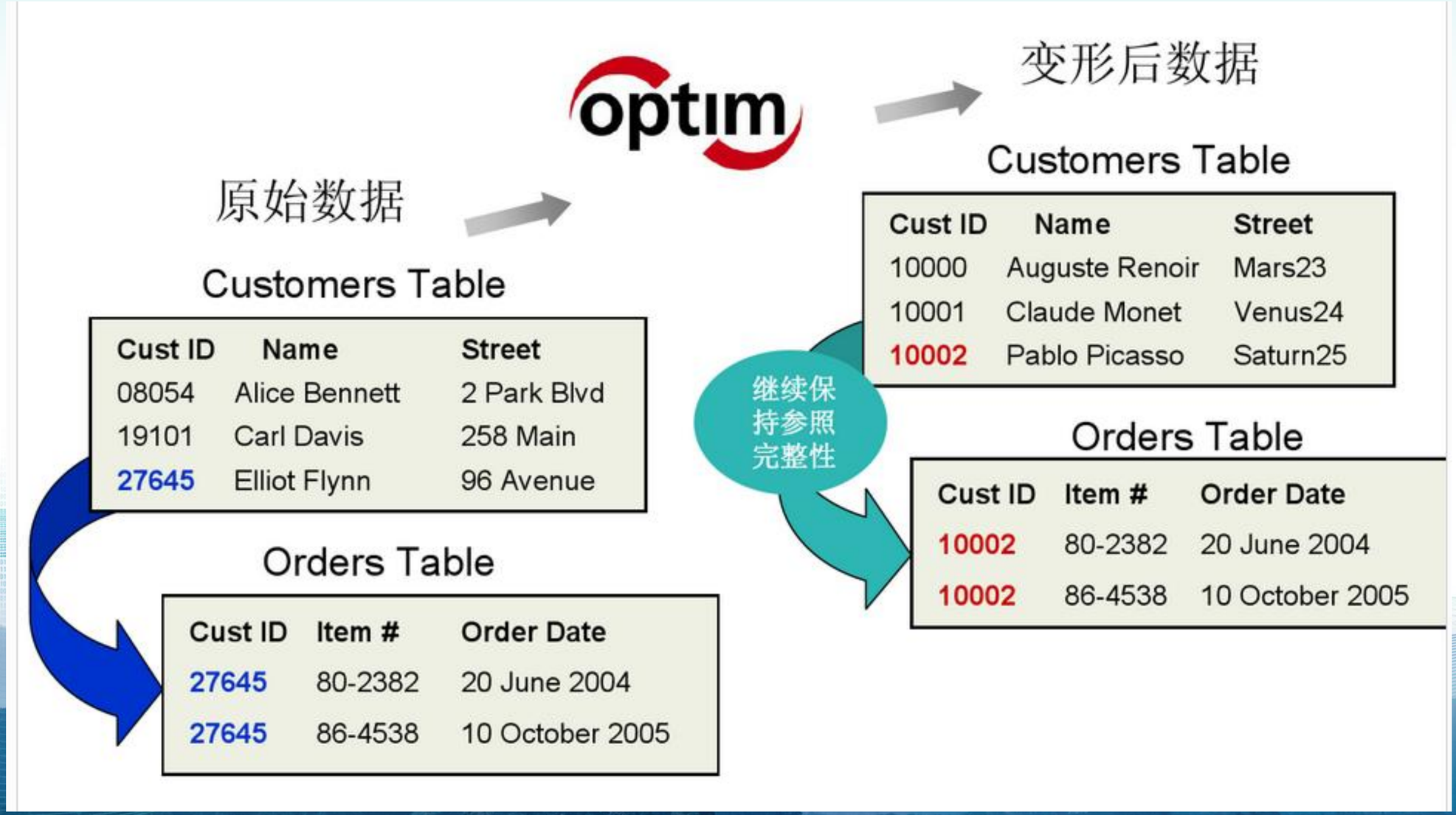
数 据 脱 敏 商 用 解 决 方 案

- IBM InfoSphere Optim数据脱敏
- Oracle数据脱敏
- Informatica数据脱敏
- 苹果的差分隐私保护

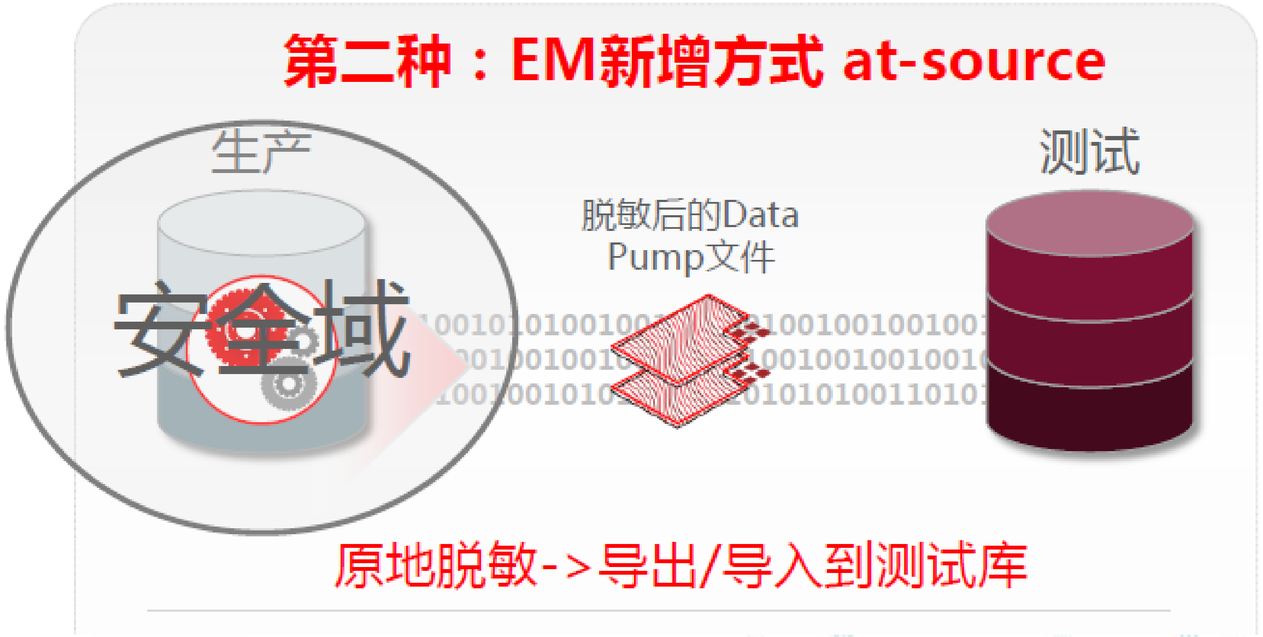
数据脱敏：IBM InfoSphere Optim



数据脱敏：IBM InfoSphere Optim



数据脱敏:Oracle



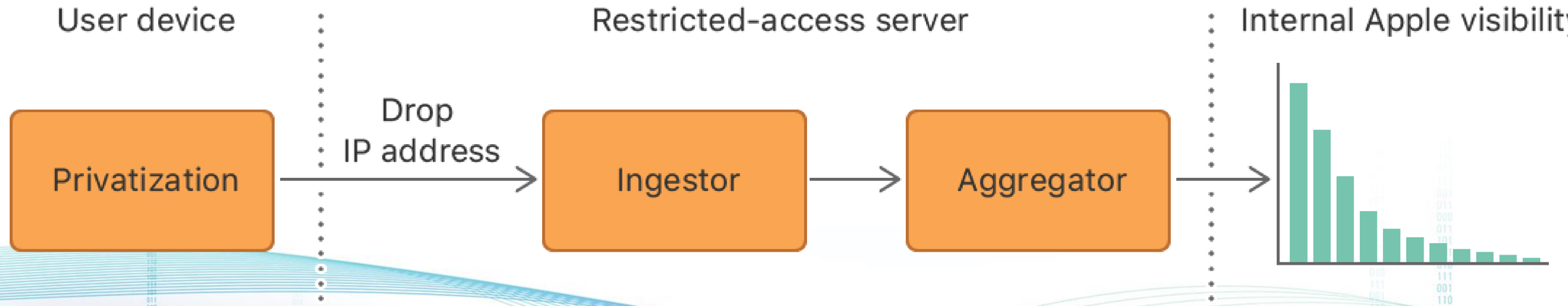
- 多种掩码技术
  - 混合掩码、基于条件的掩码、可重复掩码、打乱、加密、随机化等

数 据 脱 敏 : I n f o r m a t i c a

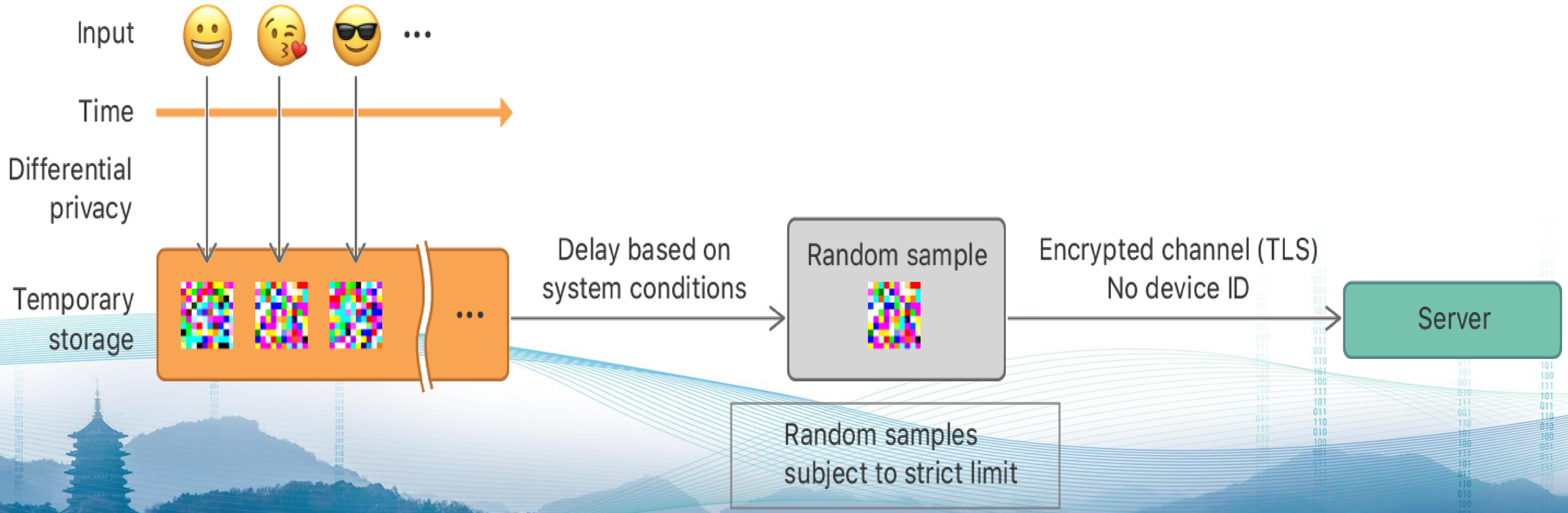
- 多种脱敏技术
  - 打乱编码ID、替换名称、常量替换、信用卡掩码技术

数据脱敏：苹果差分隐私技术

差分隐私，通过 laplace 和指数两种机制添加噪声，目标是做数据挖掘前先进行处理。苹果的方案，是在手机本地加入噪声后再上传，一般统计的是输入法的新词汇，表情包的使用状况，运动相关数据等。



数据脱敏:苹果差分隐私技术Privatization



THANK YOU

谢　　谢　　观　　看