



西安电子科技大学
XIDIAN UNIVERSITY



计算机科学与技术学院
School of Computer Science and Technology

第五章 大数据处理及安全隐私

彭延国

ygpeng@xidian.edu.cn





- 大数据特性：
 - Volume、Velocity、Variety、Veracity、Value
- 带来的困境
 - 用户端难以存储和处理
- 解决途径
 - 将数据加密后存储在云端



flickr



foursquare™

M 阿里云邮
qiye.aliyun.com

NETFLIX

EC EasyChair
The conference system

Expedia®



lyft



经济

高扩展性

不间断访问

现有的防御机制不足以应对危机！

公有云

攻陷！

交易！

破坏！

.....



西安电子科技大学
XIDIAN UNIVERSITY

在密文数据上保持明文的可搜索语义

§5.1 数据安全检索





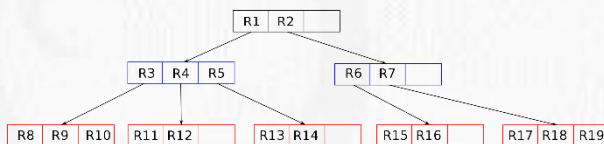
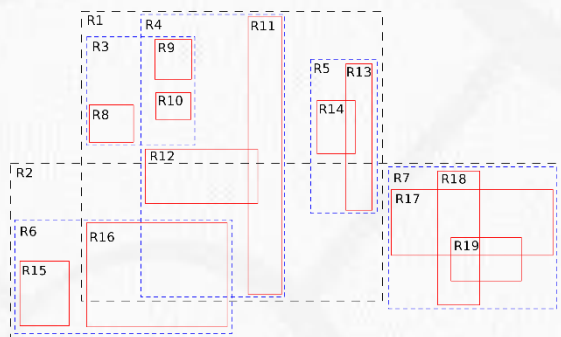
- 检索(Search, Retrieval)^[1]: 特指信息检索, 是从**大规模**非结构化数据的集合中找出满足用户信息需求资料的过程。
- 操作对象:
 - 通常是文本(文档)、数值、复杂数据(时空数据、高维数据、图像、声音等)
- 存储地点:
 - 本地计算机、存储设备
 - 云端存储设备**
- 核心任务:
 - 满足用户信息需求



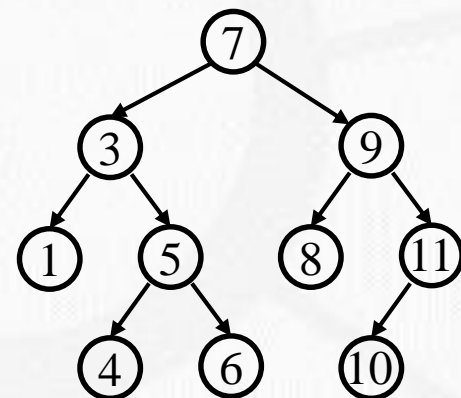
检索是用户信息需求的基石



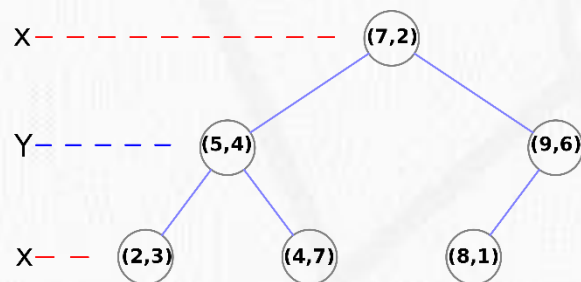
- 索引(Index)^[1]: 是一种数据结构, 目的是提高数据检索操作的速度, 代价是增加了维护索引数据结构的存储空间。
- 索引的类型:
 - 树形索引(Tree-based Index)
 - 倒排文档
- 树形索引的核心
 - 数据具有偏序关系
- 密文上的偏序关系
 - 保序加密技术



R树



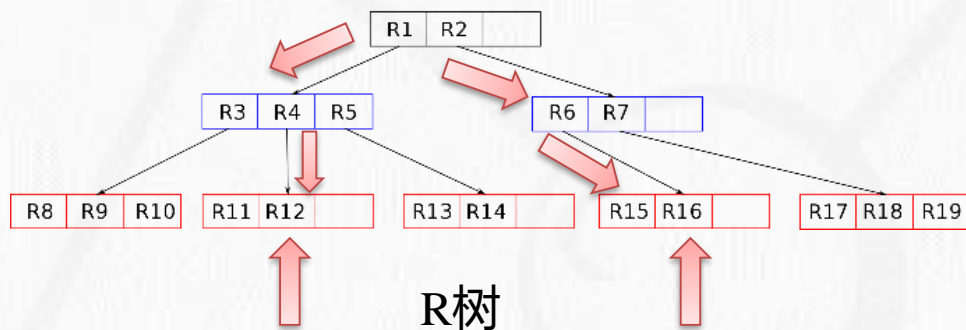
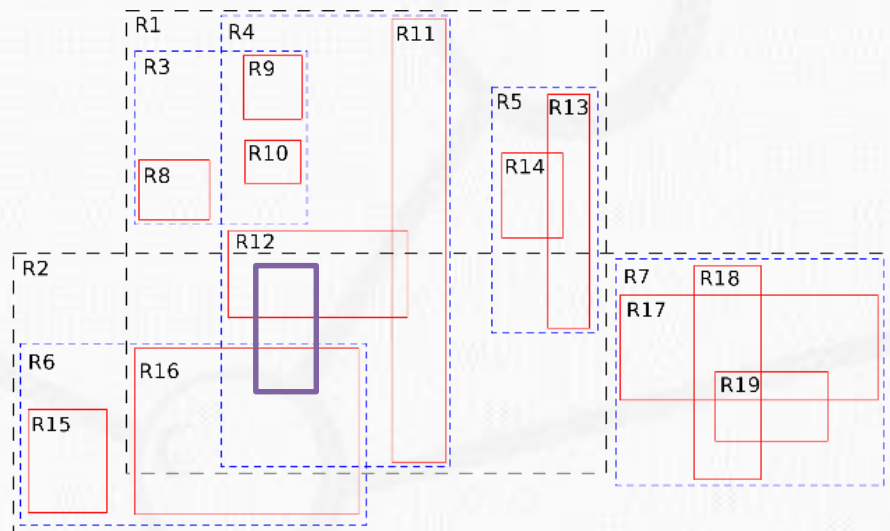
二叉树



KD树

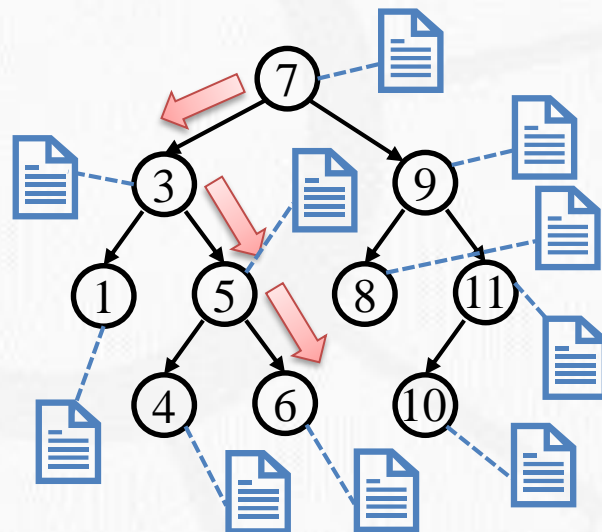


§5.1 数据安全检索 - 基于索引的查找过程



R树

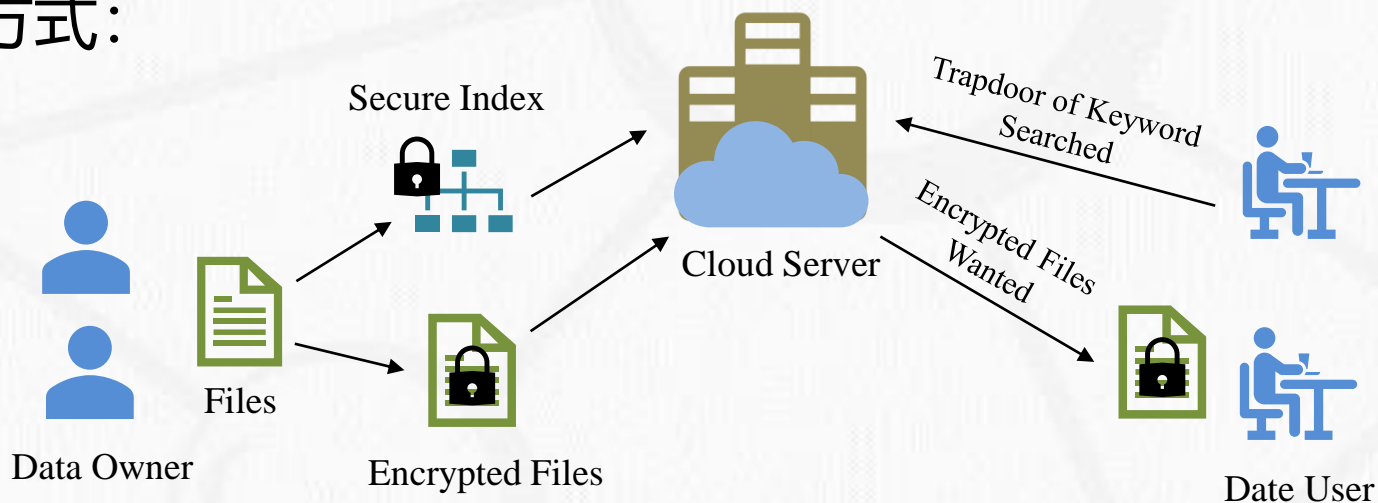
查询: 6



二叉树



- 可搜索加密(Searchable encryption, SE) ^[1]: 允许用户将数据以一种能够保证隐私的形式外包给第三方, 同时保证用户通过特定途径能够在外包的数据上进行检索。
- 保证隐私的方式:
 - 加密
- 第三方:
 - 公有云
- 特定途径:
 - 使用检索令牌





- 根据对象不同
 - 关键词、一维数据、多维数据(关系型数据、时空数据)、高维数据(图像特征)
- 根据采用的密码技术不同
 - 对称可搜索加密、非对称可搜索加密、匿名可搜索技术
- 核心工具
 - 保序加密、现代密码学工具、代理重加密、同态加密、安全多方计算等



西安电子科技大学
XIDIAN UNIVERSITY

偏序关系、索引

§5.1.1 保序加密

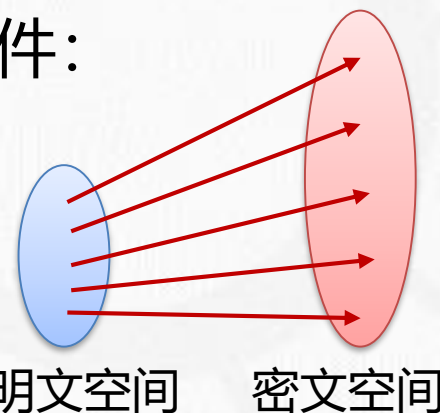




§5.1.1 保序加密 - 保序加密

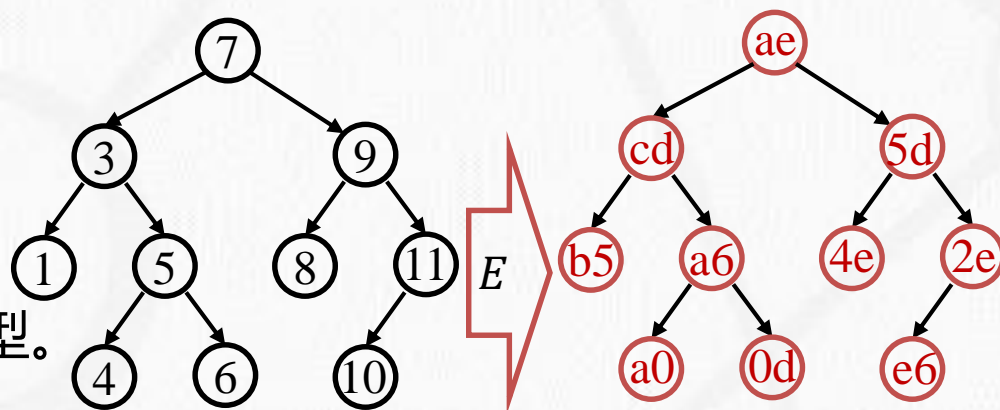
- 保序加密(Order-Preserving Encryption, OPE): 给定一个明文, $m \in M$ 、加密方案 $E_{sk}(m): \{sk, m\} \rightarrow c \in C$ 和密文 C , E 是一个保序加密方案, 当且仅当满足以下条件:

- 正确性: $D_{sk}(E_{sk}(m)) = m$ 。
- 保序性: $m_0 \leq m_1$ 当且仅当 $c_0 \leq c_1$ 。



- 安全性需求:

- IND-CPA不再适用;
- ROPF、IND-OCPA安全模型。

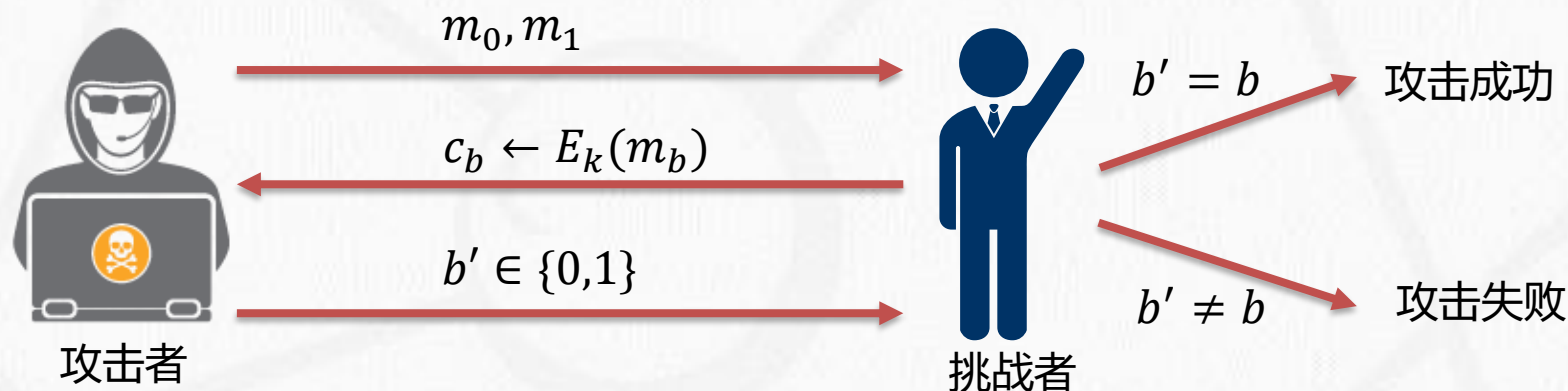




- 不可区分性(Indistinguishability)

- 攻击者(Adversary)对两个明文 m_0 和 m_1 进行挑战。挑战者(Challenger)随机选择其中的一个明文用给定的加密方案加密得到密文 c_b ，并发送给攻击者。攻击者猜测 $b' \in \{0,1\}$ ，若满足下式则称该加密方案是不可区分安全的：

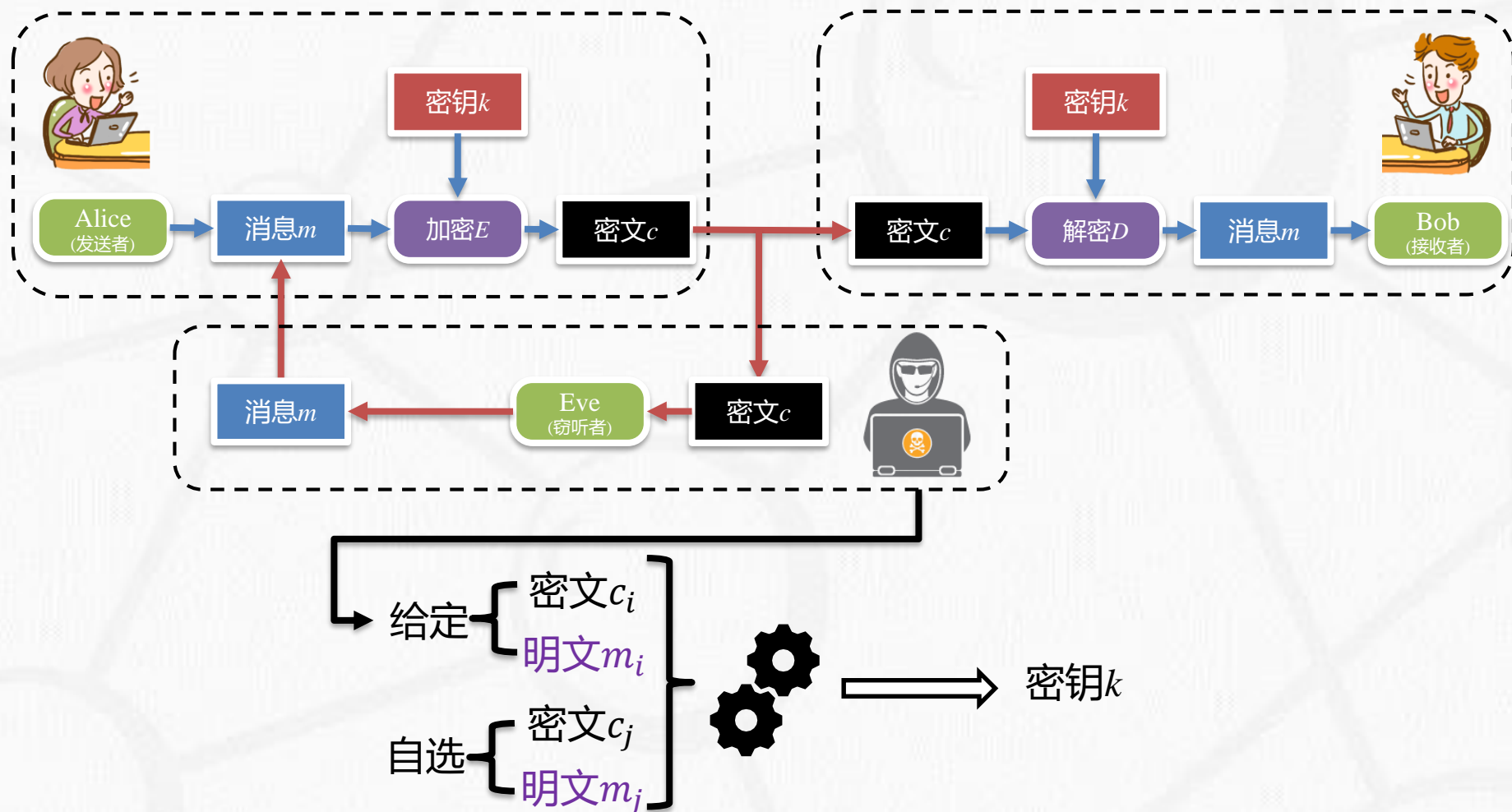
$$- \left| \Pr[b = b'] - \frac{1}{2} \right| \leq \text{negl}(n).$$





- 选择明文攻击(CPA: Chosen-Plaintext Attack)

- 攻击者自主选择明文 m_i , 并获得对应的密文 c_i





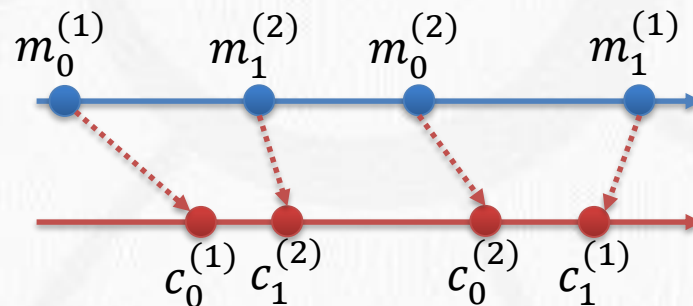
为什么IND-CPA不适用于保序加密？

– IND-CPA的核心思想：

- 任给两个明文 m_0 和 m_1 ，攻击者 \mathcal{A} 进行加密挑战，得到 c_0 和 c_1 ，随机返回一个密文 $c_b, b \in \{0,1\}$ 。 \mathcal{A} 进行猜测 $b' \in \{0,1\}$ ，如果 $b' = b$ ，则攻击成功。



单轮的CPA攻击



基于CPA攻击攻破保序加密

- 显然的，经过两轮的CPA攻击，可以明显猜测出第一轮 b' 。
- 因此，从基本理论上可以推导出：**任何保序加密方案均不可能达到IND-CPA安全。**



- IND-OCPA安全(Indistinguishability under ordered chosen plaintext attack):
 - 给定一个协议 Π , 如果对于所有的概率多项式时间的攻击者 \mathcal{A} 存在可忽略函数 $negl$, 满足
$$\text{Adv}_{\Pi}^{\text{IND-OCPA}}(\mathcal{A}) = |\Pr[\text{Exp}_{\Pi}^{\text{IND-OCPA}} = b] - \frac{1}{2}| \leq negl$$
 - 则保序加密方案 Π 是IND-OCPA安全的。也称为理想安全的。
- 通俗来讲, 理想安全性要求密文除去明文顺序信息而不泄露明文的其他任何信息。
- 此外, 保序加密的安全模型还有:
 - IND-FAOCPA等



- Liu等^[1]提出的保序加密方案
 - 密钥: a, b
 - 加密函数: $E(m) = am + b + n$, 其中 n 为噪声
- 一种潜在的CPA攻击:
 - 例子: $a = 10, b = 3, n \in [1, 3]$
 - 选择的明文序列: $m_1 = 0, m_2 = 5, m_3 = 13, m_4 = 21, m_5 = 102$
 - 获得的密文序列: $c_1 = 4, c_2 = 55, c_3 = 136, c_4 = 214, c_5 = 1026$
 - 攻击原理: $c_i - c_1 = am_i + n_i - n_1$

$$\frac{c_i - c_1}{m_i + 1} \leq a \leq \frac{c_i - c_1}{m_i - 1}$$



- Liu等^[1]提出的保序加密方案
 - 密钥: a, b
 - 加密函数: $E(m) = am + b + n$, 其中 n 为噪声
- 一种潜在的CPA攻击:
 - 攻击原理: $\frac{c_i - c_1}{m_i + 1} \leq a \leq \frac{c_i - c_1}{m_i - 1}$
 - 攻击结果: 推断出 $a = 10$
- 进一步可以推断出全部明文信息。
- 无法达到IND-OCPA安全。

序号	m	c	a
1	0	4	-
2	5	55	[9,12]
3	13	136	[10,11]
4	21	214	[10,11]
5	102	1026	[10,10]



- 理想安全的保序加密技术:

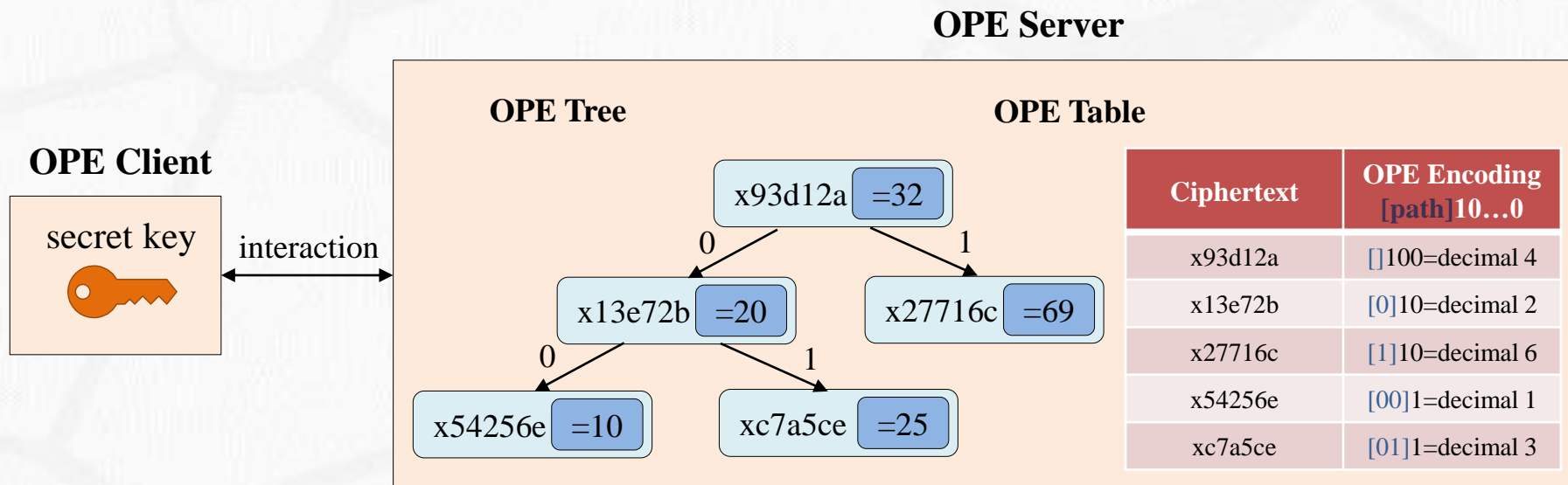
明文:

69	32	20	10	25
----	----	----	----	----

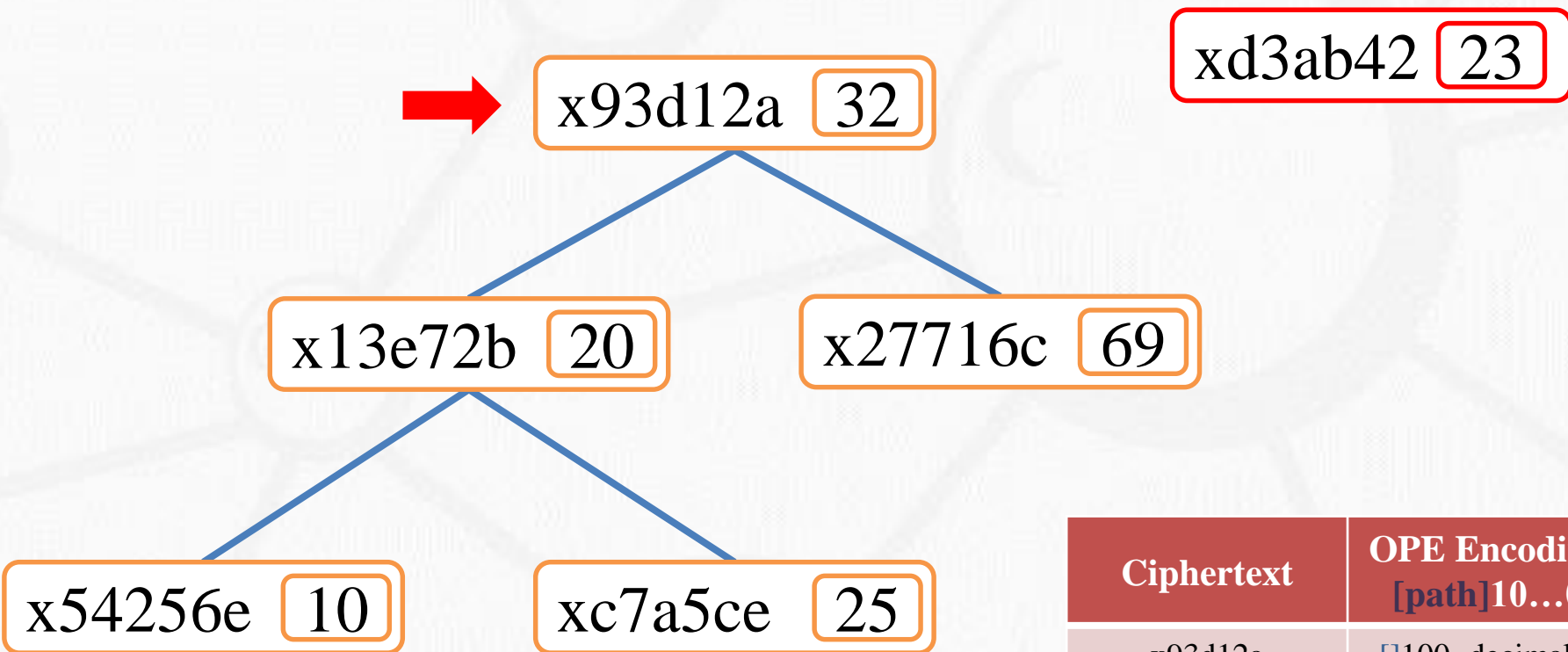
密文:

5	4	2	1	3
---	---	---	---	---

- Popa等^[1]提出的mOPE



[1] R Popa, F Li, N Zeldovich. An ideal-security protocol for order-preserving encoding[C]. IEEE S&P. 2013: 463-477.



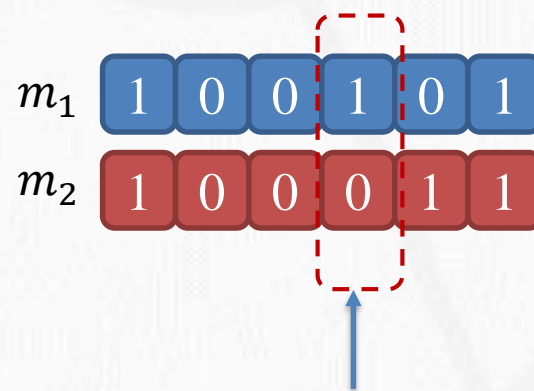
Ciphertext	OPE Encoding [path]10...0
x93d12a	[]100=decimal 4
x13e72b	[0]10=decimal 2
x27716c	[1]10=decimal 6
x54256e	[00]1=decimal 1
xc7a5ce	[01]1=decimal 3



- 现象:
 - 仍然可以进行任意明文的比较。
 - 需要在用户和服务端之间进行交互。
- 存在的问题:
 - 任意攻击者可以对密文进行比较, 造成密文数据的滥用。
 - 存在交互, 效率很低。
 - 需要密文映射表才能解密。

• 如何解决问题?

- 除了顺序外, 再多泄露一点点信息。



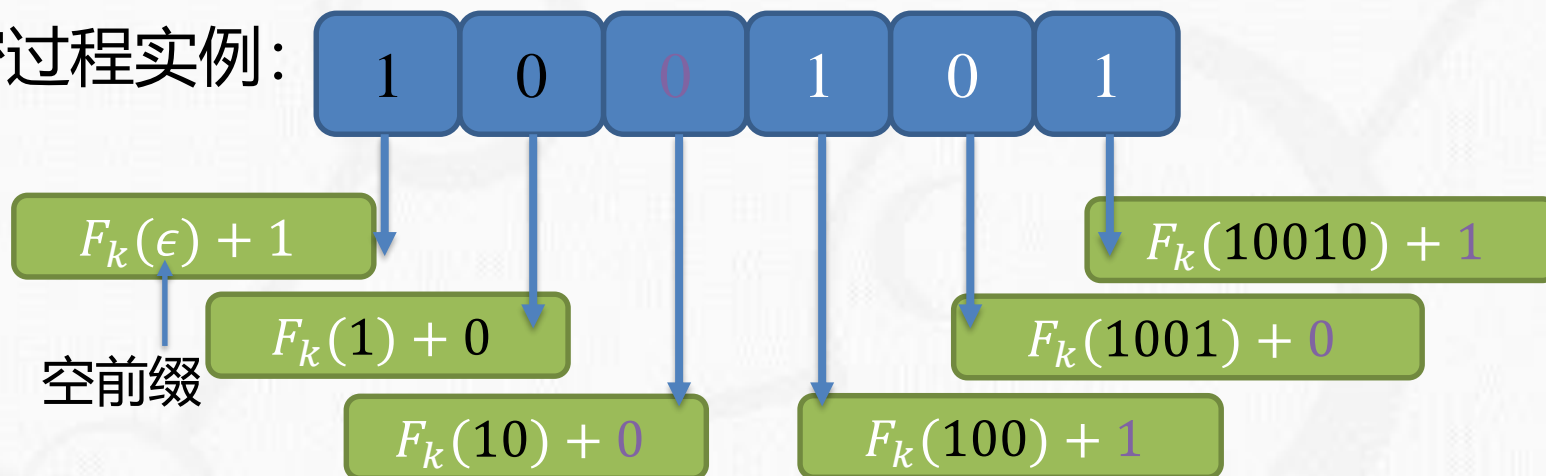
$\text{ind}_{\text{diff}}(m_1, m_2)$: 第一个不相同的比特位



- 揭序加密(Order-revealing encryption): 一种有陷门的顺序可见加密方案。
 - 伪随机函数 (Pseudorandom function, PRF) F 的密钥 k , $m = \{b_1, b_2, \dots, b_n\}$ 。
 - 加密: $c = \{u_i\}_{i=1}^n, u_i = F(k, (i, b_1 b_2 \dots b_{i-1} || 0^{n-i})) + b_i \bmod M, M \geq 3$ 。
 - 比较: c 和 c' 的大小。
 1. 找到第一个不相同的 u_i 和 u_i' 。
 2. 如果没找到, 返回 $c_i = c_i'$ 。
 3. 如果两个密文第 i 个位不同: 如果 $u_i' = u_i + 1 \bmod M$, 则 $c_i < c_i'$; 否则 $c_i > c_i'$ 。
 - 解密: $b_i = F(k, (i, b_1 b_2 \dots b_{i-1} || 0^{n-i})) \oplus u_i$



加密过程实例:



密文比较实例:



顺序决定位

$F_k(\epsilon) + 1$ $F_k(1) + 0$ $F_k(10) + 0$

相同的前缀=相同的密文

$F_k(\epsilon) + 1$ $F_k(1) + 0$ $F_k(10) + 0$

$F_k(100) + 1$

第一个不同的
加密块

$F_k(100) + 0$

$F_k(1001) + 0$ $F_k(10010) + 1$

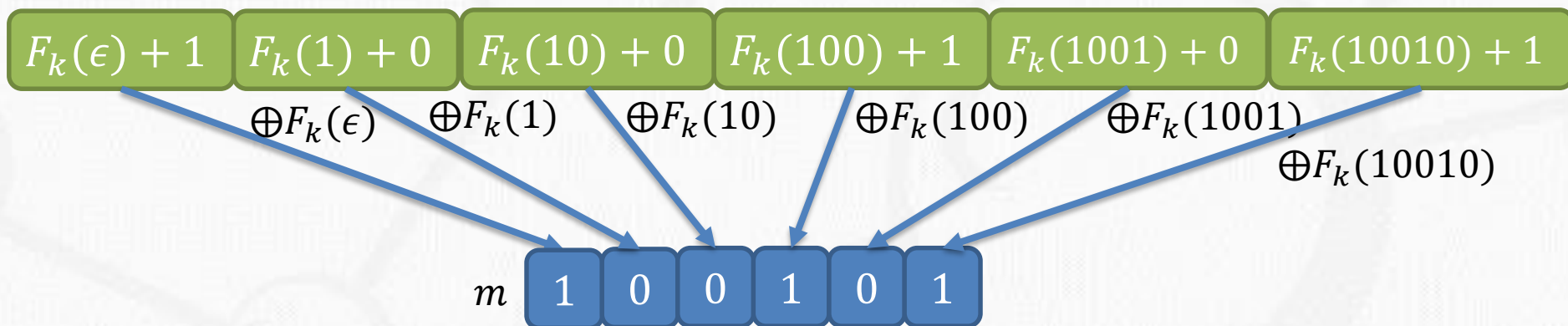
不同的前缀=隐藏真正的值

$F_k(1000) + 1$ $F_k(10001) + 1$

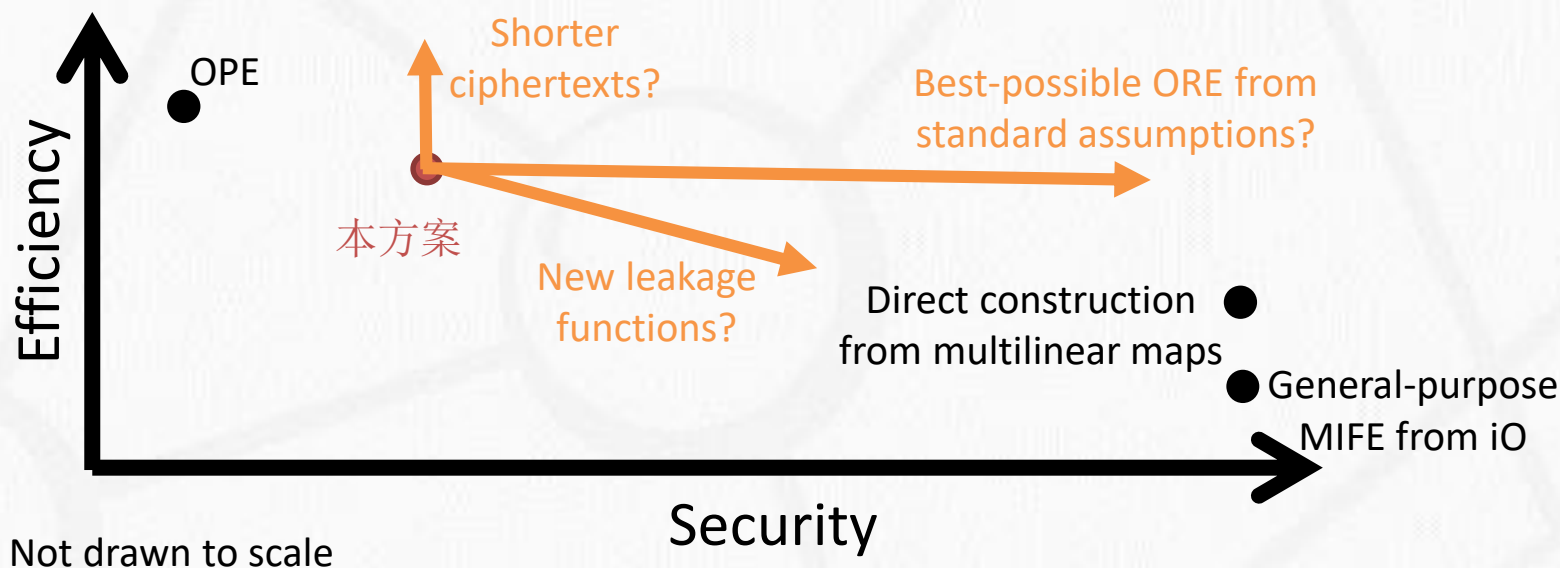




- 解密过程实例: F, k



- 未来研究方向:





西安电子科技大学
XIDIAN UNIVERSITY

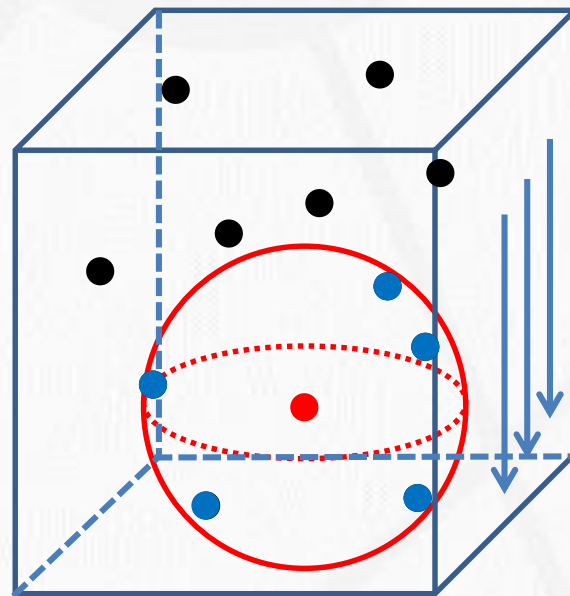
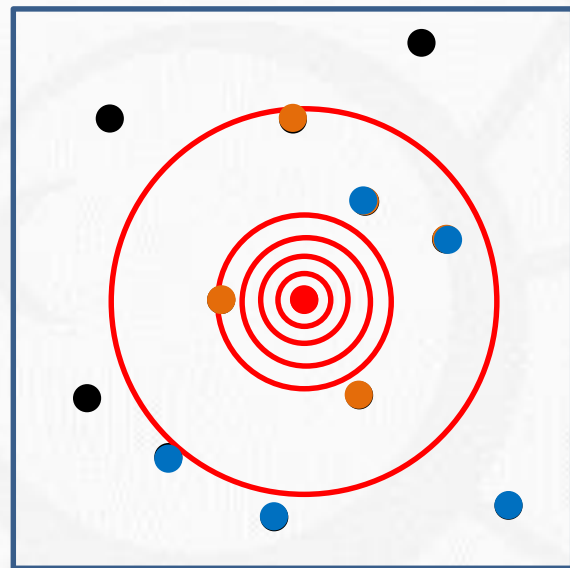
多维数据、高维数据

§5.1.2 多维数据安全检索





- 最近邻(NN)查询
 - 寻找与目标点距离最近的点。
- k 最近邻(k NN)
 - 寻找与目标点距离最近的 k 个点。
- 近似最近邻(ANN)
 - k NN的相似解。



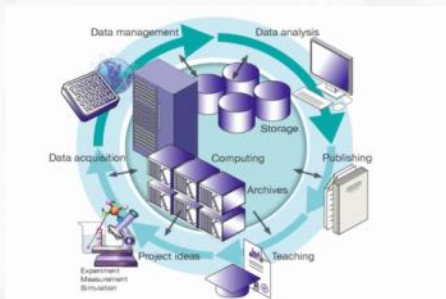


- Beyer et al. (1999)
 - “维度灾难”，在高维空间中，任意两点间距离极其接近。

- 近似最近邻（ANN）应用场景



Computer Vision Data Management



GIS

- 近似最近邻的本质
 - 效率和准确性的折衷。

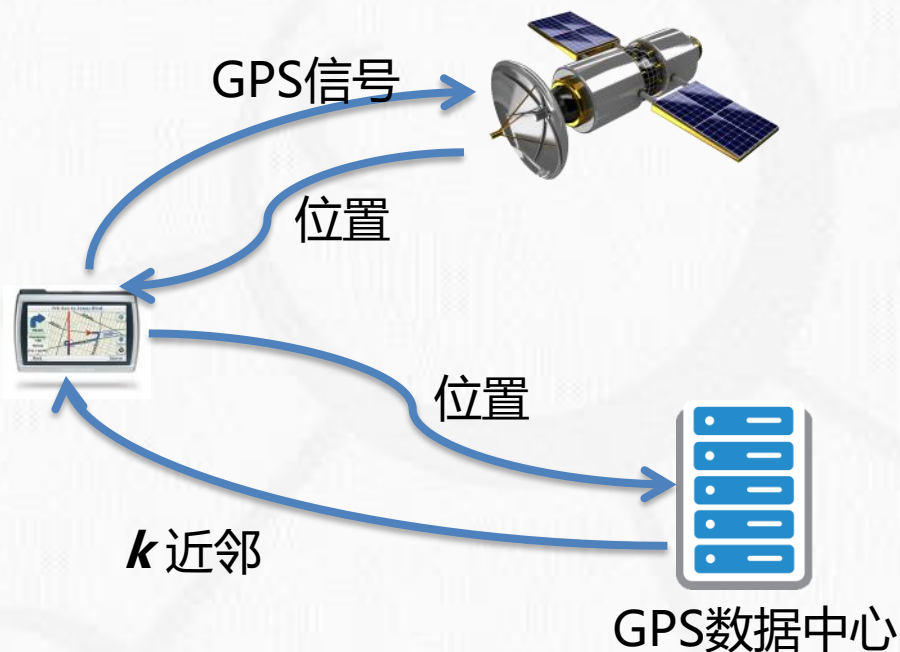


- 地理信息系统(GIS: Geographic Information System)

- 位置隐私性
- 需要安全传送

- 目标:

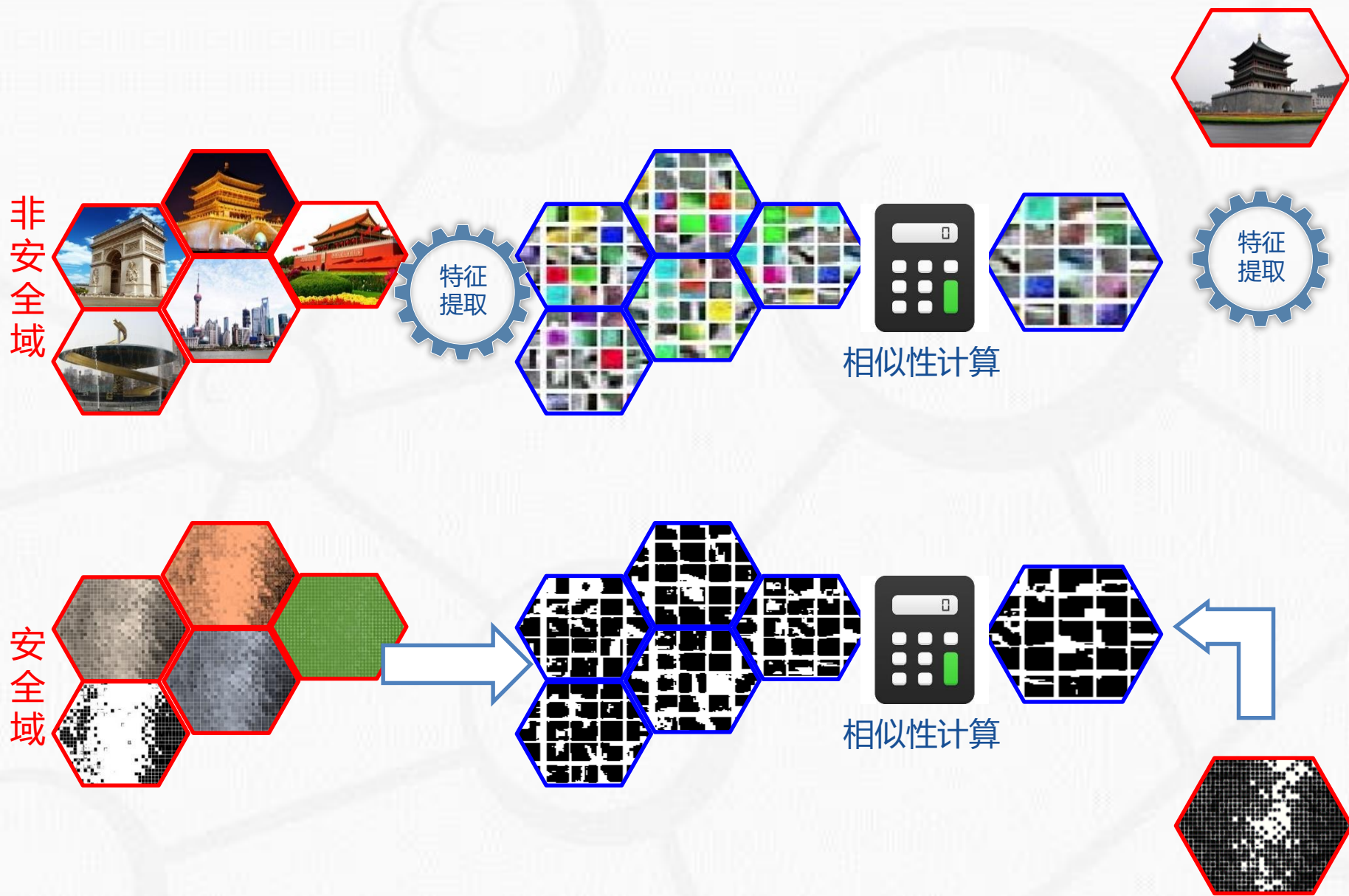
- 保护用户位置不被泄露



用户的查询 (位置) 被保护!

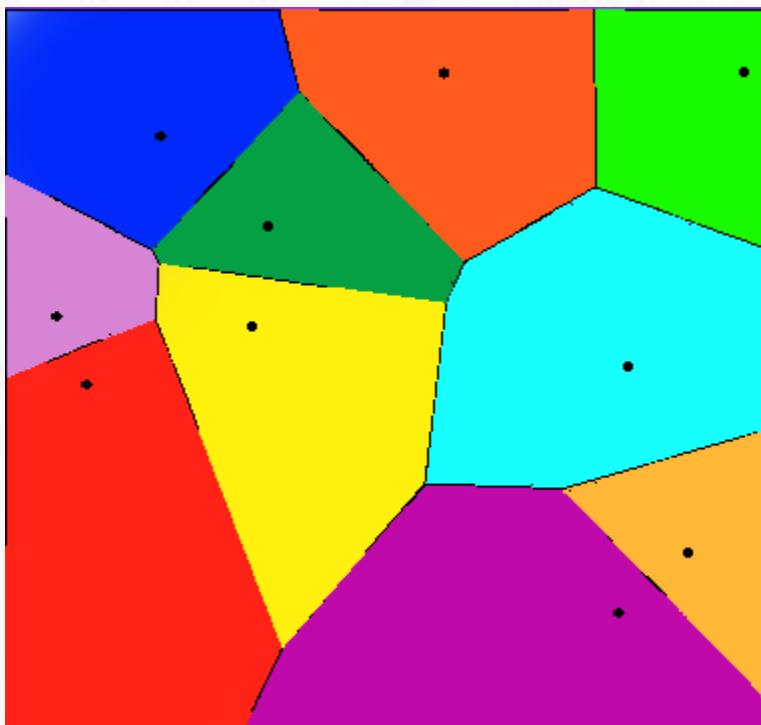


§5.1.2 多维数据安全检索 - 图像检索的例子





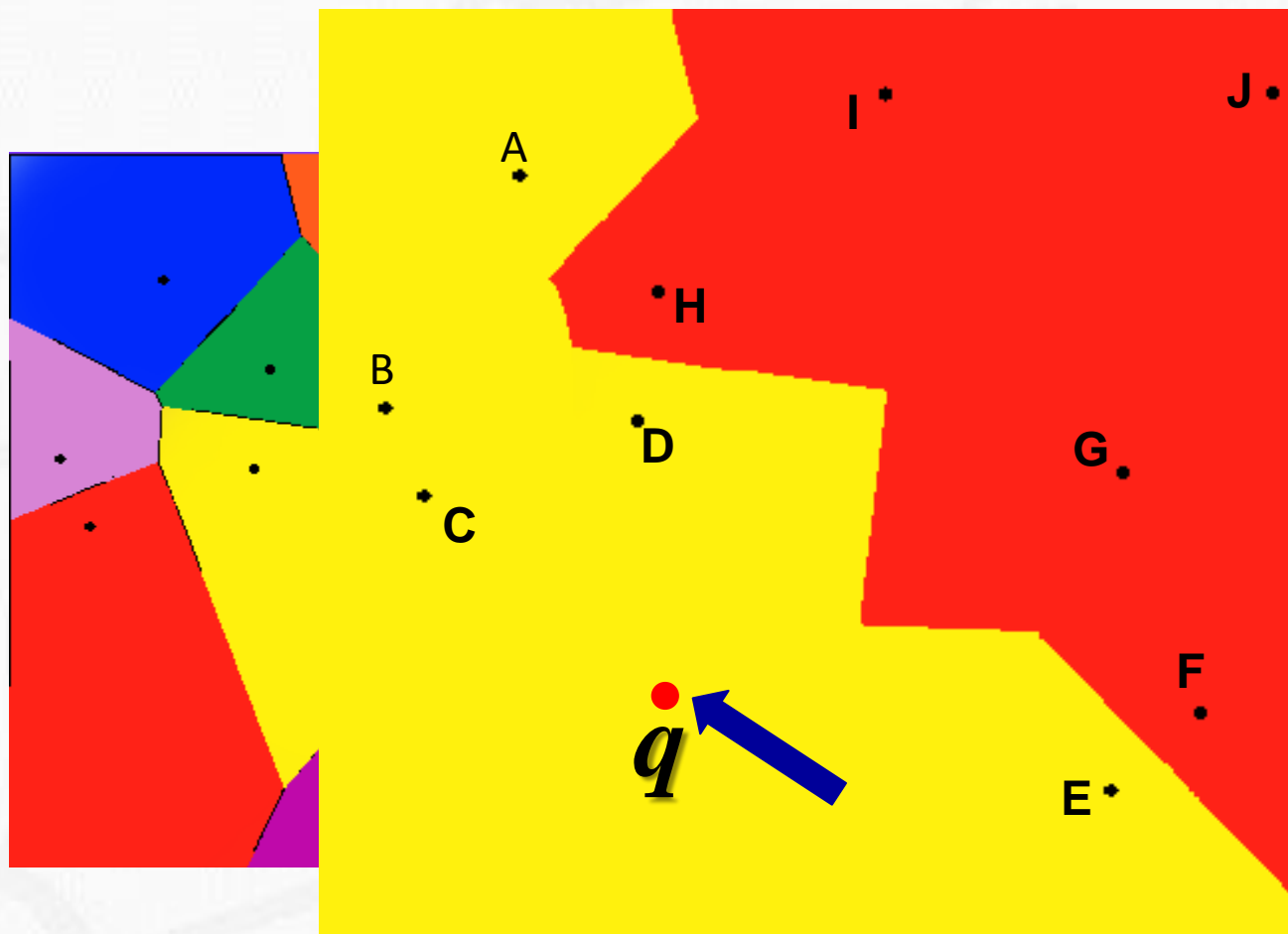
- 泰森多边形
 - 将空间划分为多个区域的典型方法。



- 每个区域内包含离中心点最近的所有点。



- 简单的近邻检索思路

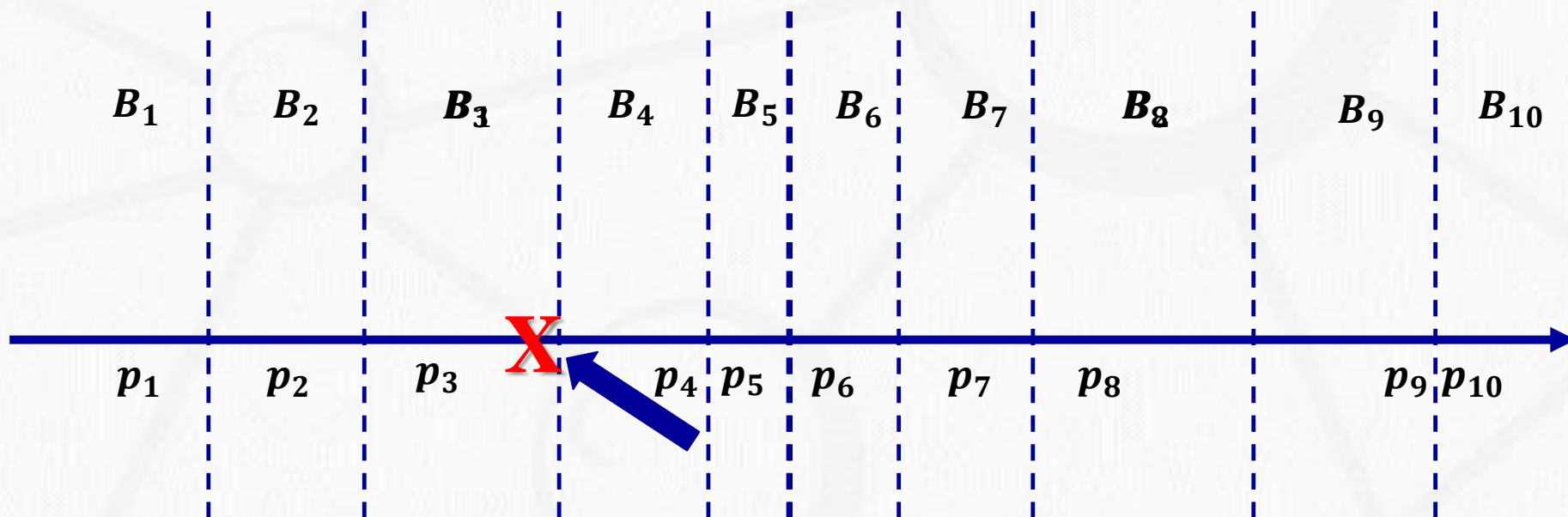




- 问题:
 - 如何确定一个查询落在哪个区域?
 - 如何将一个图划分成若干区域?
 - 如何确保每个区域有相同的点的数量?
- 主要结果:
 - 针对一维数据的分区方法、
 - 三个针对二维数据的分区方法
 - Square Grid(SG)
 - Minimum Space Grid(MinSG)
 - Minimum Max partition(MinMax)



- 需求:
 - 垂直分界线是没有意义的;
 - 不存在临界区域的概念。



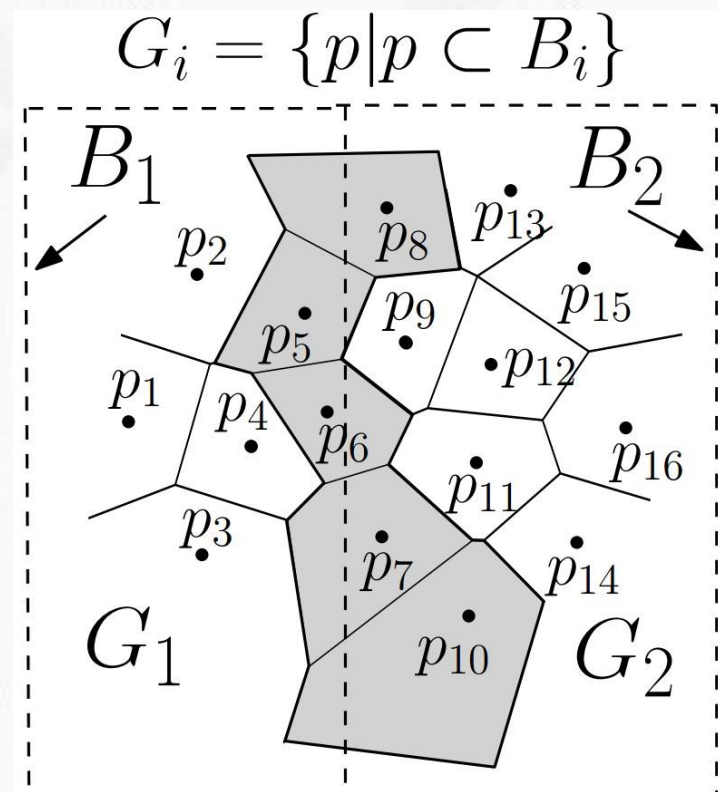


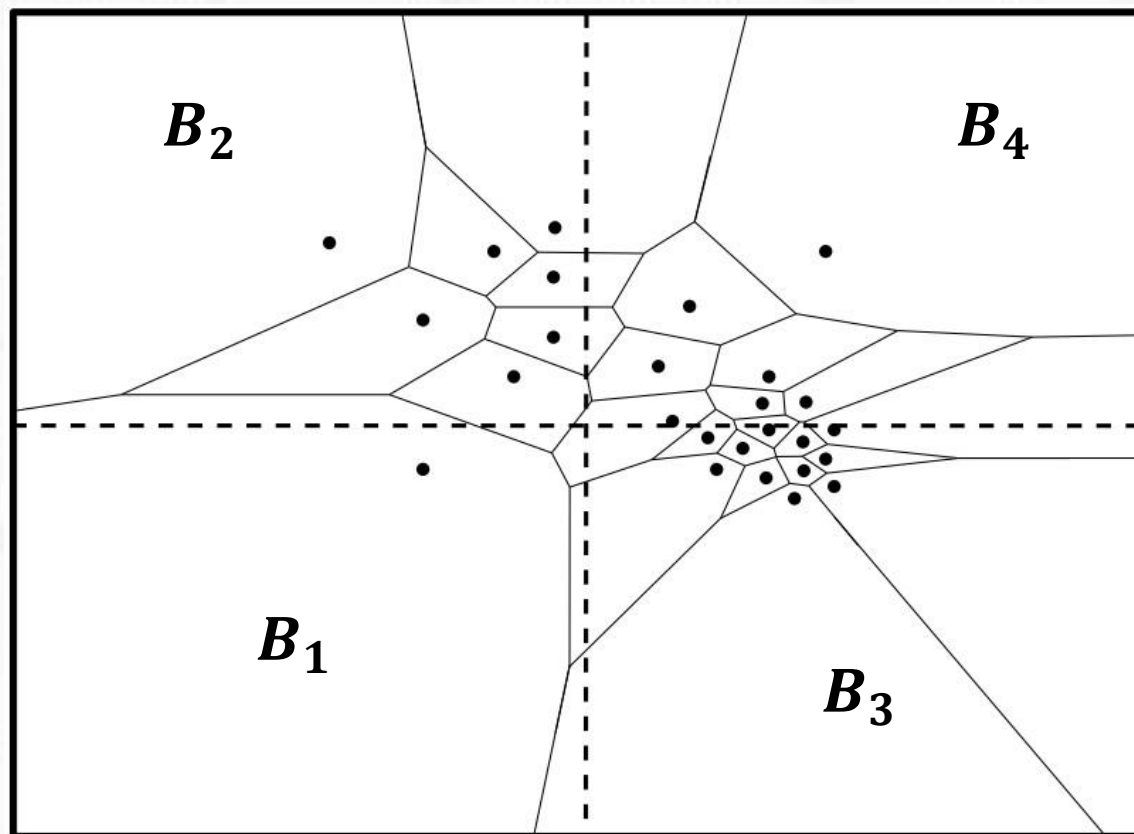
- 定义:

- D 是二维坐标数据集
- 每一个分区 (i) 对应每一个网格 B , (ii) 包含所有与当前网格 B 有重叠的泰森多边形
- $G_i = \{p | p \in B_i\}$

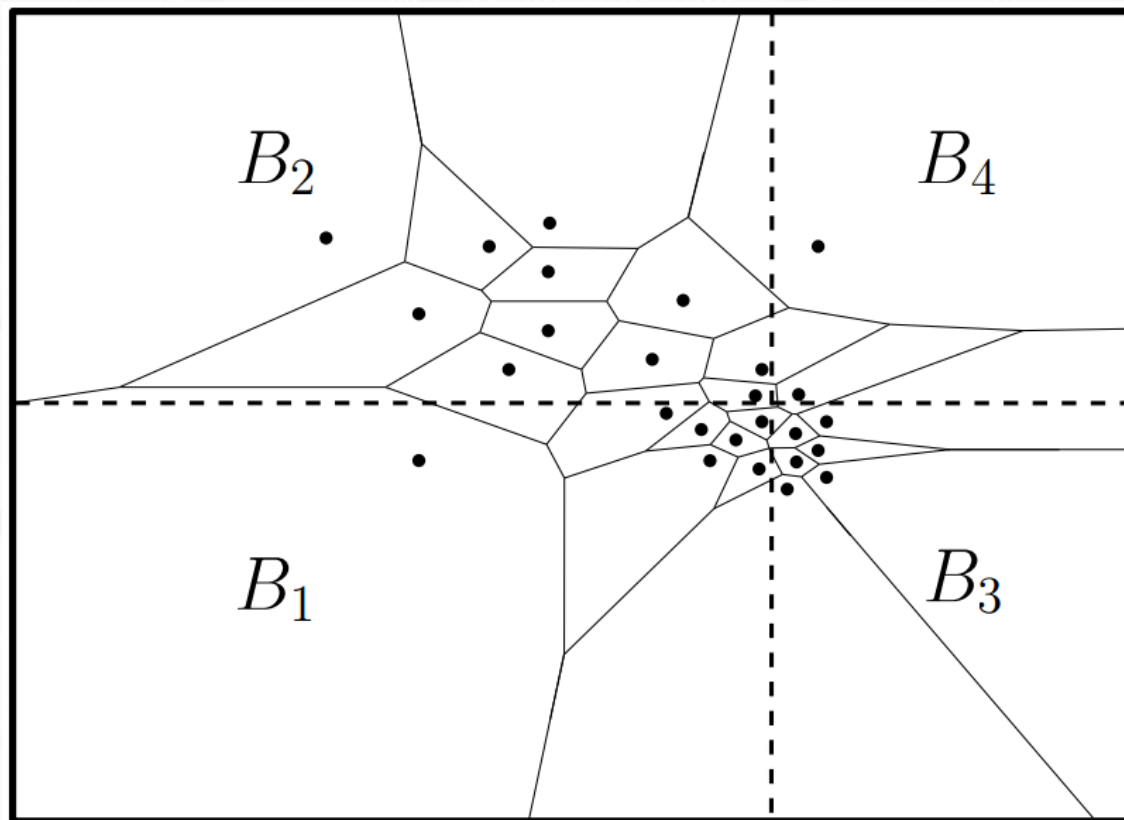
- Sample:

- $G_1 = \{p_1, p_2, p_3, p_4, p_5, p_6, p_7, p_8, p_{10}\}$
- $G_2 = \{p_6, p_7, \dots, p_{16}\}$

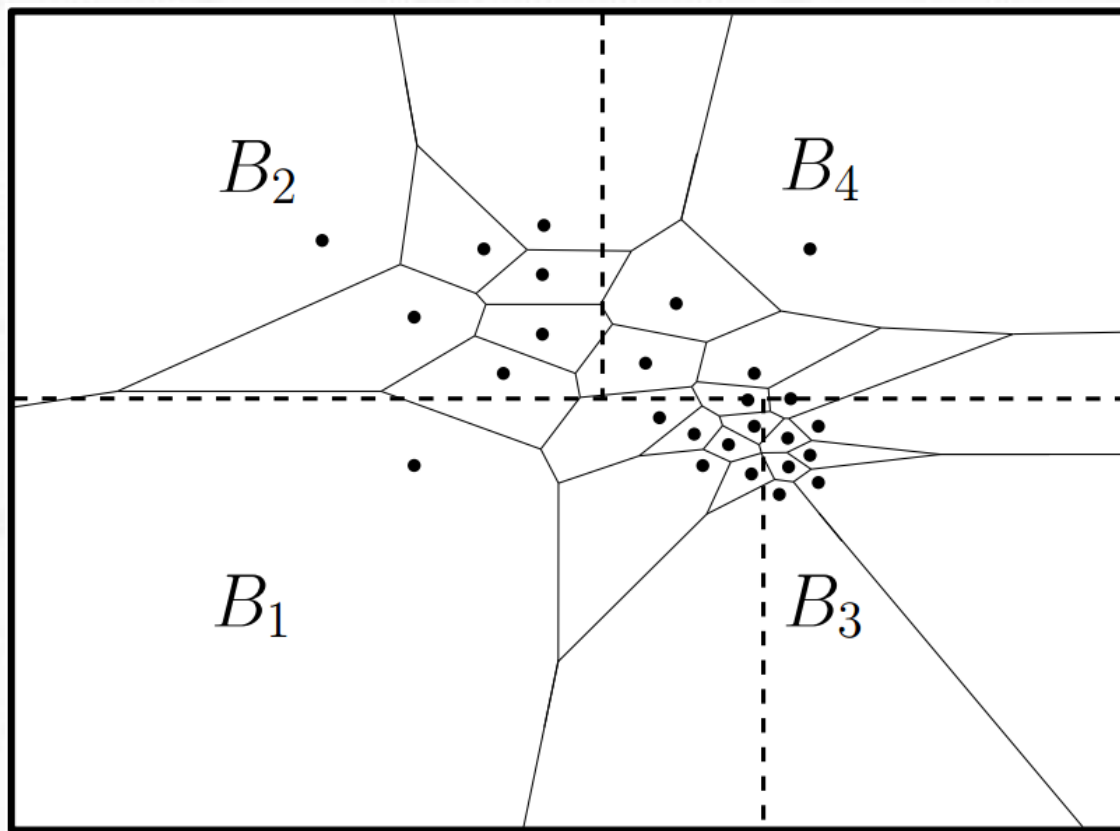




- 问题: $|B_3| \gg |B_i|$, 其中 $i = 1, 2, 4$ 。



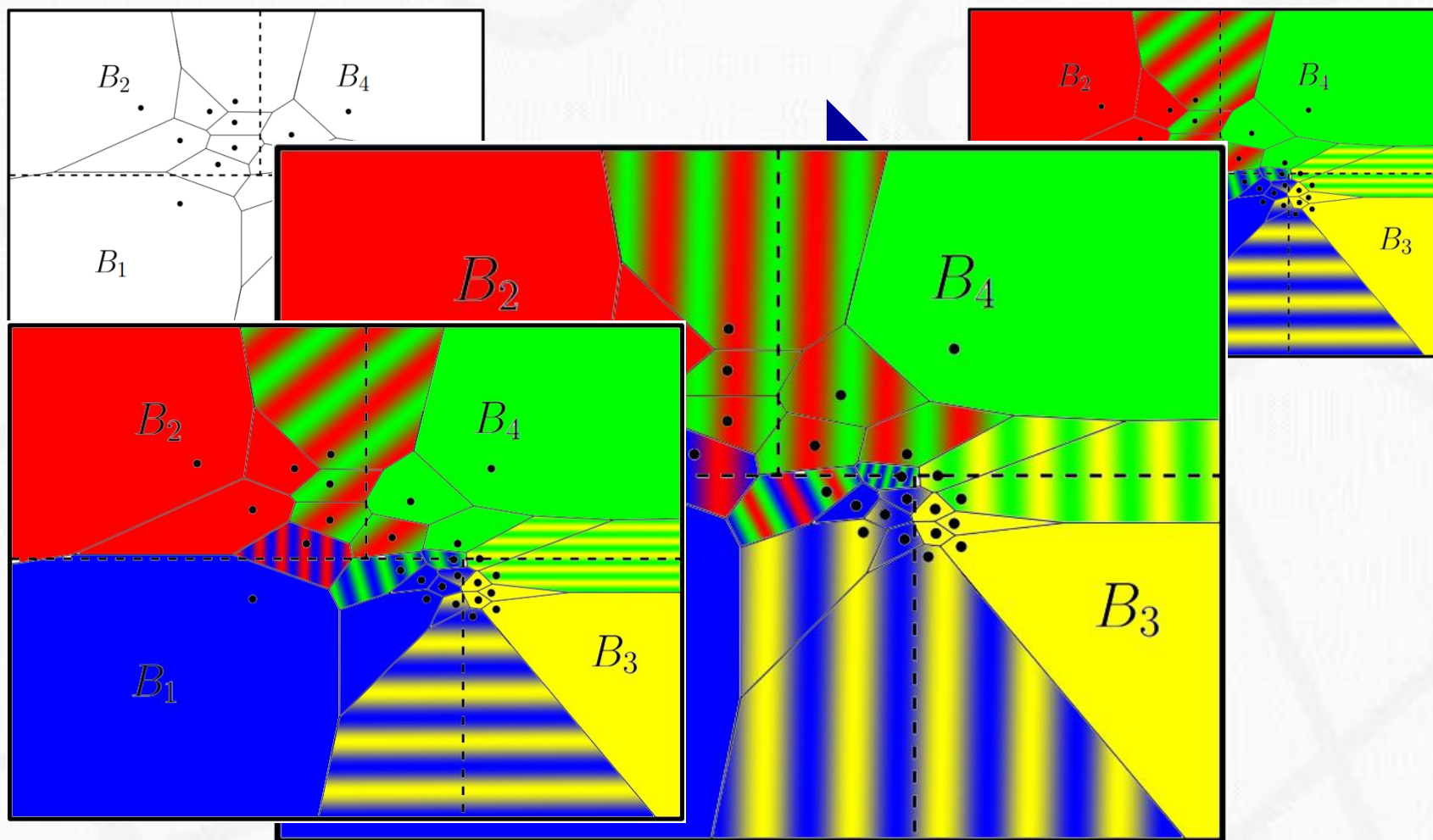
- 问题: $|B_4| \ll |B_i|$, 其中 $i = 1, 2, 3$ 。



- 问题：对于 SG、MinSG和 MinMax, $|B_i| \neq |B_j|$, 其中 $i \neq j$.



- B_4 包含12个点，其它的区域也扩展到12个点。





二维以上数据是否适用?



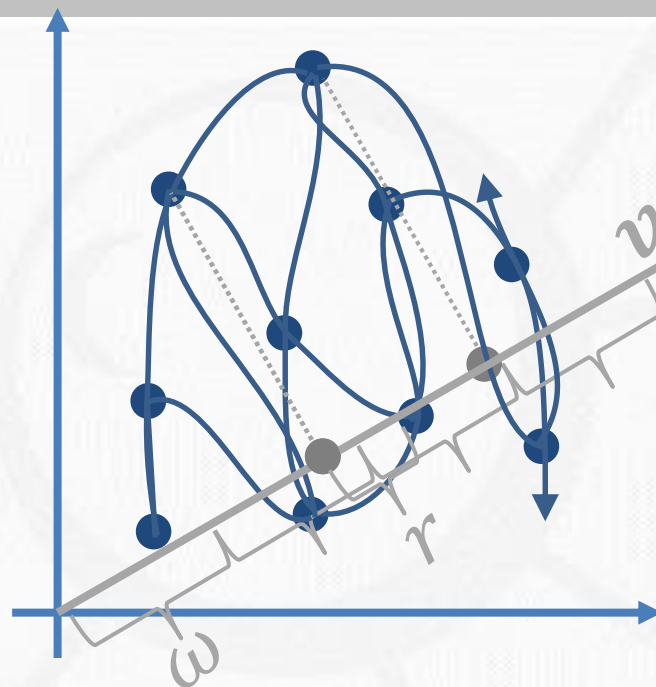
- Datar et al. (2004)
 - 多、高维数据的排序问题;
 - 点到数的转换。

$$h(v) = \left\lfloor \frac{\theta \cdot v + r}{\omega} \right\rfloor$$

- (R_1, R_2, P_1, P_2) -敏感哈希

- 对于 $\|q - v\| \leq R_1$, $\Pr[h(v) = h(q)] \geq P_1$;
- 对于 $\|q - v\| \geq R_2 = cR_1$, $\Pr[h(v) = h(q)] \leq P_2$;

$$P_1 = \int_0^\omega f(t) \left(1 - \frac{t}{\omega}\right) dt, P_2 = \int_0^\omega \frac{1}{c} f\left(\frac{t}{c}\right) \left(1 - \frac{t}{\omega}\right) dt,$$





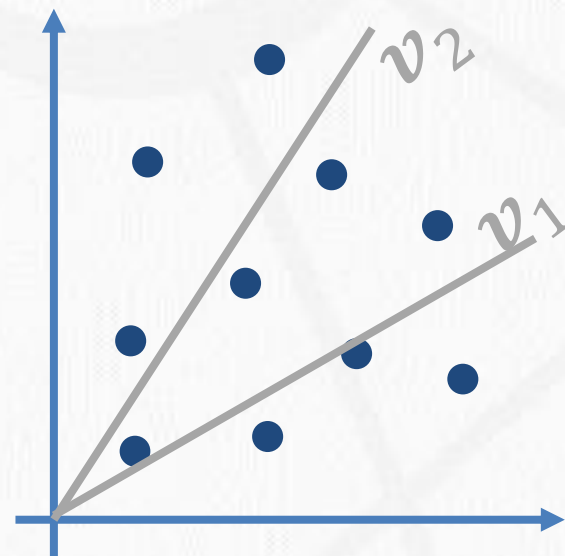
- 变种的位置敏感哈希

$$h(v) \equiv \left\lfloor \frac{\left\lceil \frac{\theta(v) + r}{\omega} \right\rceil - \min_{p \in \mathbb{D}} \{y(p)\}}{\omega} \right\rfloor$$

– $y(v) = \theta \cdot v$, 其中 $\omega = (\max_{p \in \mathbb{D}} \{y(p)\} - \min_{p \in \mathbb{D}} \{y(p)\}) \cdot 2^{-\lambda}$ 。

- 复合敏感哈希函数

– $G_m(v) = (h_1(v), h_2(v), \dots, h_m(v))$ 。



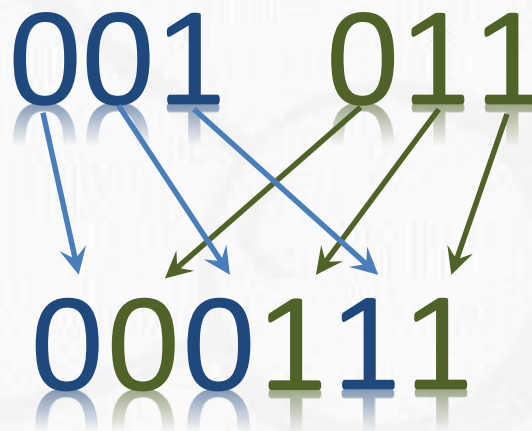


- 给定:

- $p_1 = (2.5, 2), \alpha_1 = (7, 1), \varpi_1 = 11.5$

- 可得:

- $G_m(p_1) = (1, 3)$



- 编码值:

```
1: function CODE( $p, H, \lambda, m$ )
```

```
2: → Compute the compound LSH value  $H_p = (h_1, h_2, \dots, h_m)$ ;
```

```
3:  $c = \emptyset$ ;
```

```
4: for each  $i = 1$  to  $\lambda$  do
```

```
5:     for each  $j = 1$  to  $m$  do
```

```
6:         The  $i$ -th bit of  $h_j$  is embedded into  $c$ ;
```

```
7:     end for
```

```
8: end for
```

```
9:     return  $c$ ;
```

```
10: end function
```



借助保序加密实现安全范围检索！



- CRT:
 - 结合 R*-树 和 AES 加密。

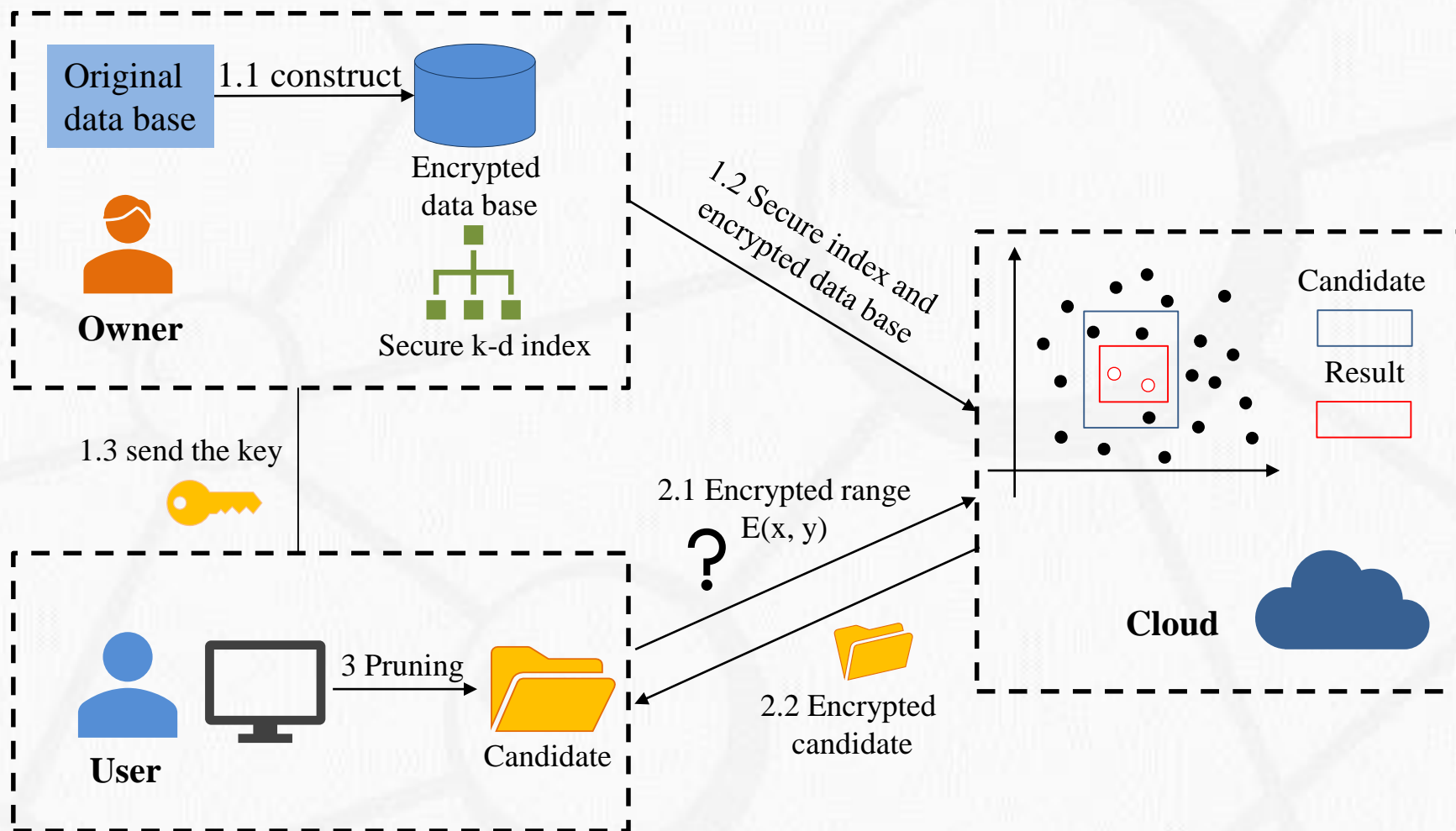
探讨方案

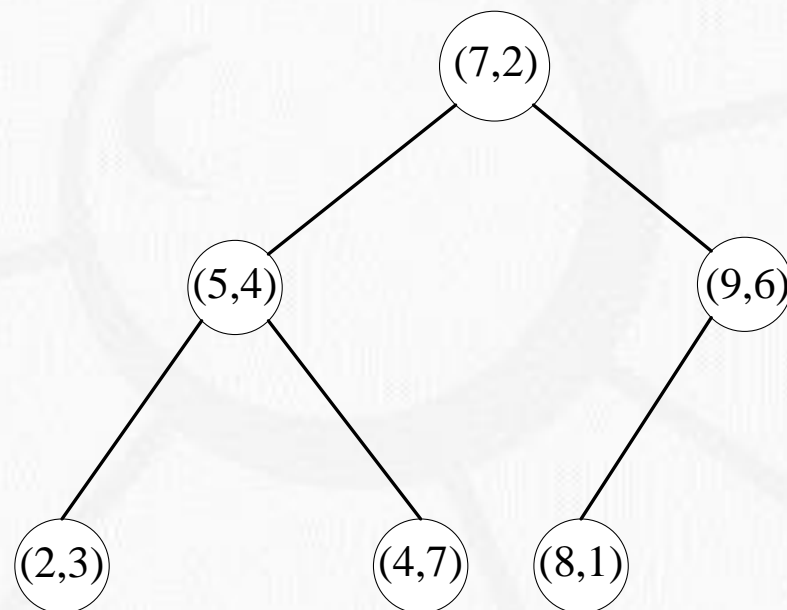
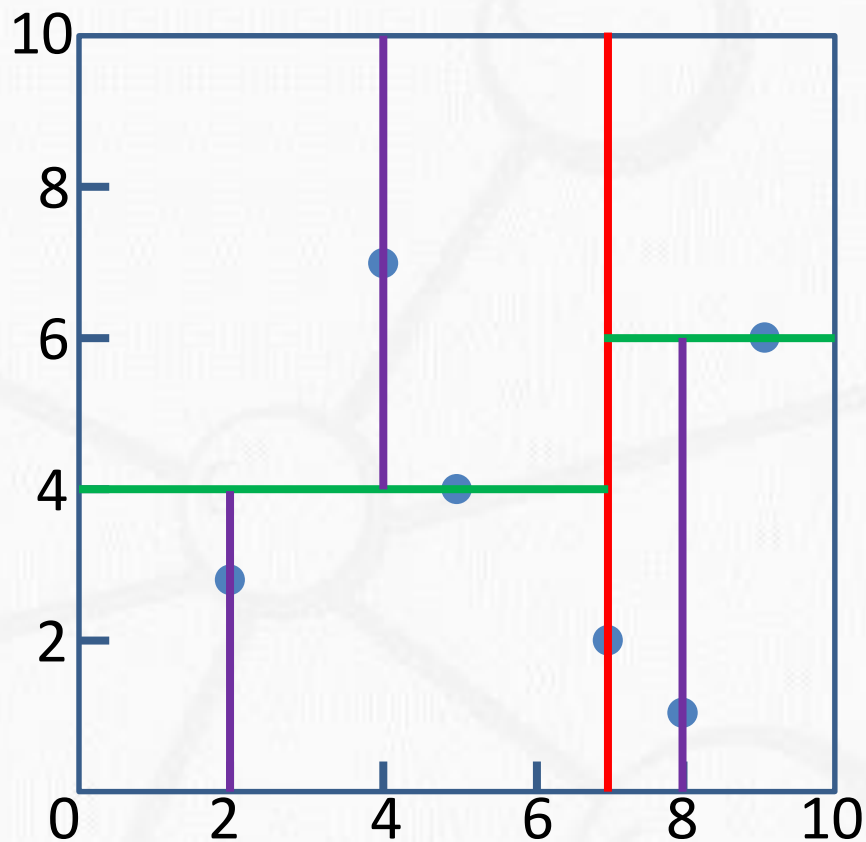
Table I: Roughly comparisons between SKD and CRT

	CRT	SKD
Index Structure	R* tree	k-d tree
Encryption Scheme	AES	CE and AES
Overhead of user end	High	Low
Round trip	Tree height	1
Communication cost	High	Low



§5.1.2 多维数据安全检索 - 方案概览

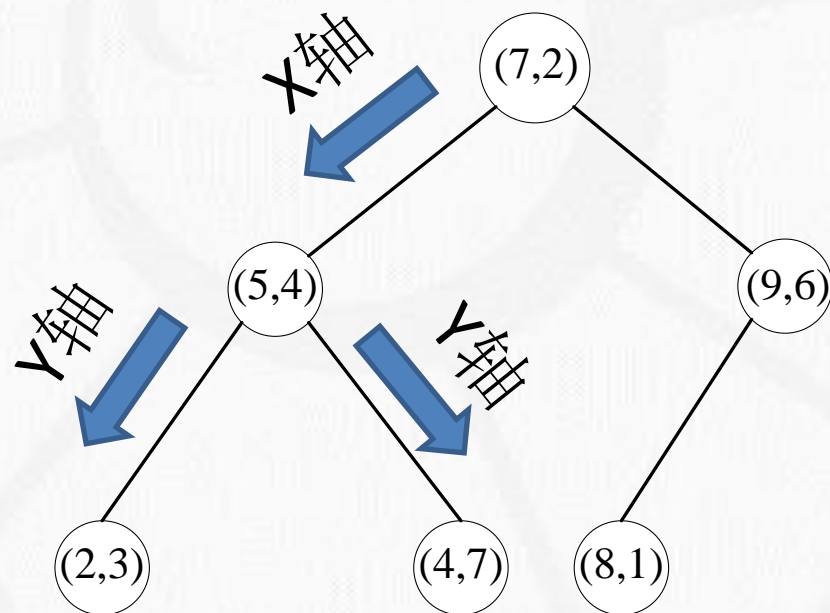
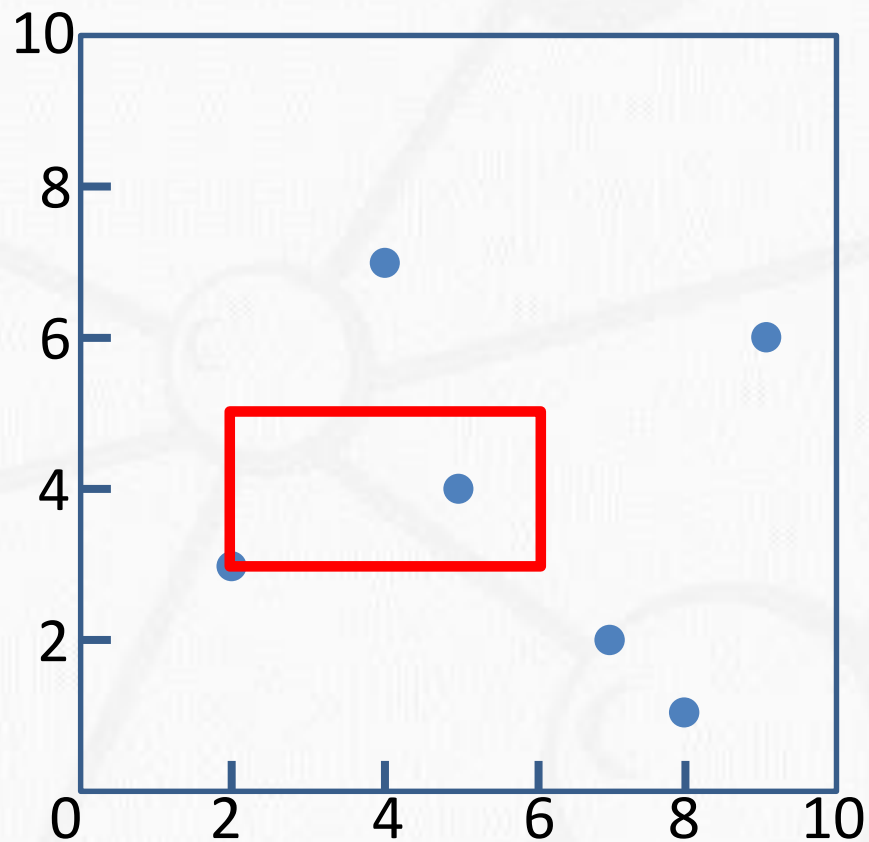




- 节点的每个分量采用保序加密(比如CE、ORE) 进行加密。



- $Q=\{(2,3),(6,5)\}$



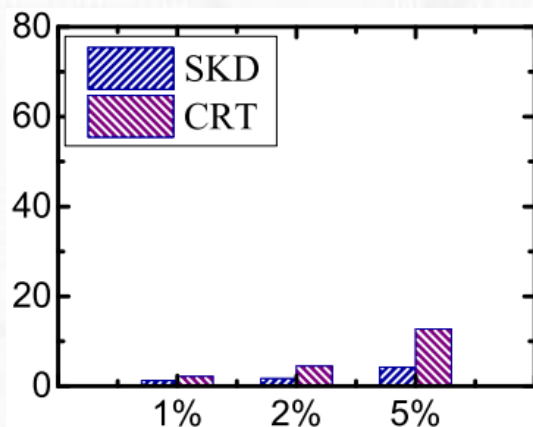


名称	范围	d	规模
TG	圣华金县	2	18,263
SF	旧金山路网	2	174,956
NA	美国北部路网	2	175,813
NE	三座城市的邮件地址	2	123,593

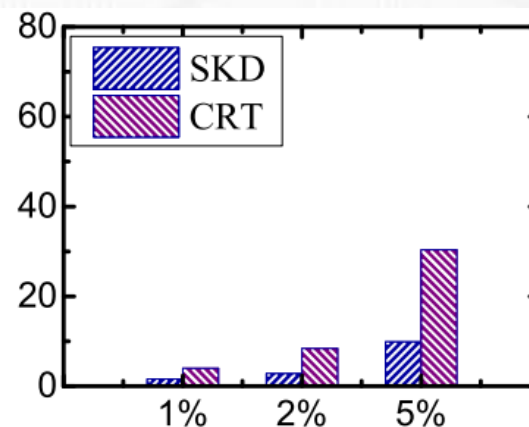
- 分别选取整个区间的1%, 2%, 5% 作为查询。



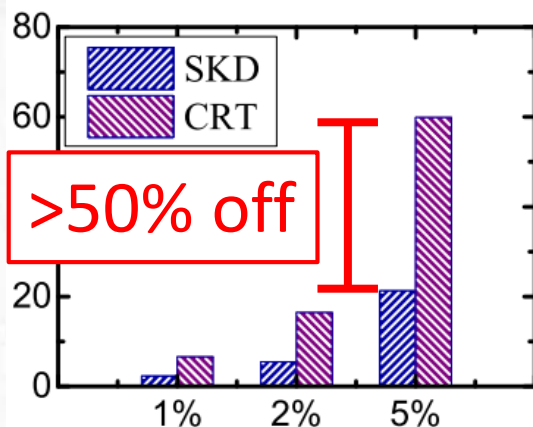
- 通信开销 (KB)



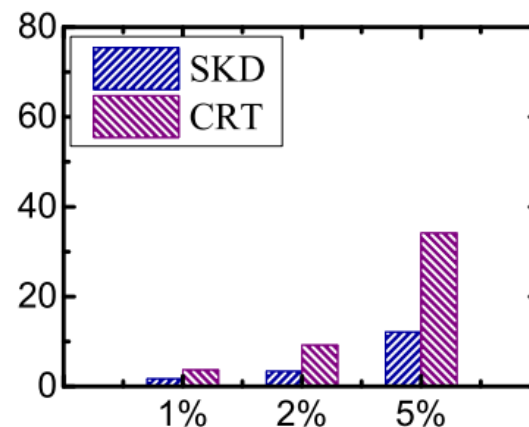
(a) Dataset TG



(b) Dataset NE



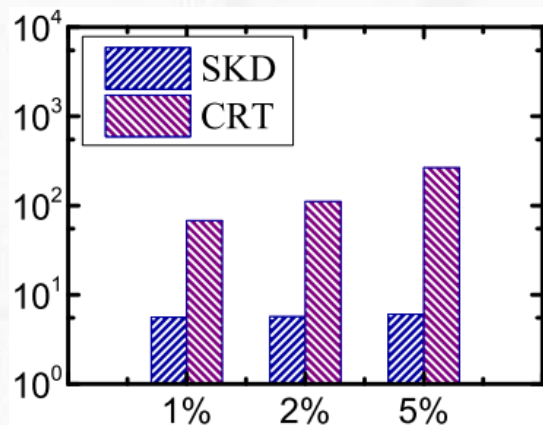
(c) Dataset SF



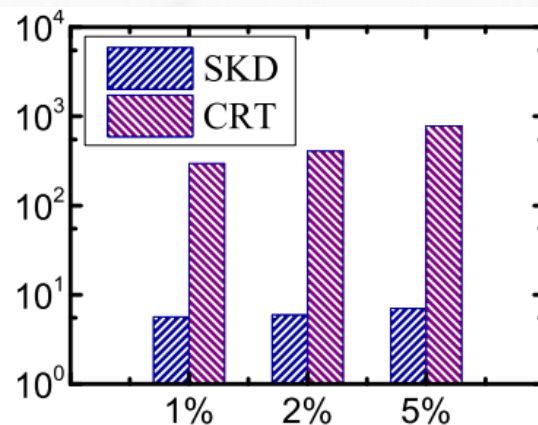
(d) Dataset NA



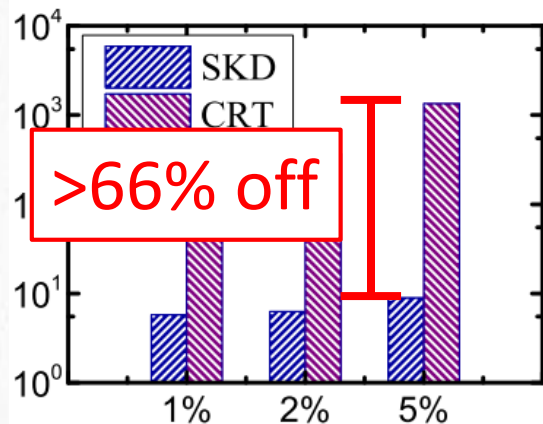
- 用户CPU时间(ms)



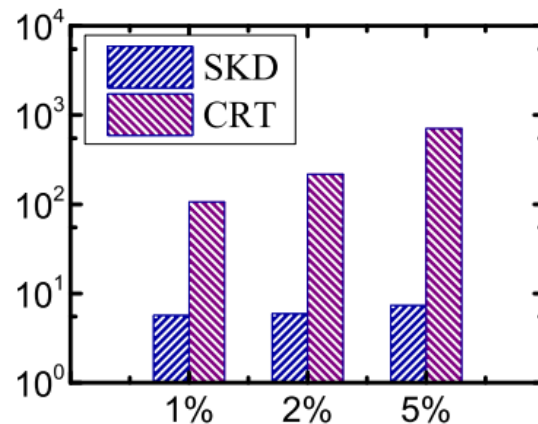
(a) Dataset TG



(b) Dataset NE



(c) Dataset SF



(d) Dataset NA



- 通信轮数

数据集	查询范围(%)	SKD	CRT
TG	1/2/5	1/1/1	6/9/21
NE	1/2/5	1/1/1	10/16/48
SF	1/2/5	1/1/1	14/30/97
NA	1/2/5	1/1/1	9/18/57

Peng Yanguo, Li Hui, Cui Jiangtao, Zhu Yixiao, Peng Changgen. An efficient range query model over encrypted outsourced data using secure k-d tree.//*In NaNA 2016: 2016 International Conference on Networking and Network Applications, 2016*, pp. 250-253. doi:10.1109/NaNA.2016.31. (EI)



- 内容回顾

- 本节介绍了保序加密的技术发展，重点讲解了四个保序加密方案：OPE、mOPE、CE和ORE，分别针对保序加密的功能、效率、安全性进行了讲解。
- 讲解了安全近邻检索、范围检索的内容。

- 掌握

- IND-OCPA和IND-CPA的区别。
- 安全近邻检索和范围检索的一般思路。



西安电子科技大学
XIDIAN UNIVERSITY



计算机科学与技术学院
School of Computer Science and Technology

Thanks!
Questions & Advices!

