1
予備戦の
加工手法

2
本戦の
加工手法

PWS Cup 2023
Team05 F.SE

大阪大学 旧藤原研
松本 知優・手島 宏貴・杉浦 一瑳
・後藤 勇芽輝・三浦 尭之

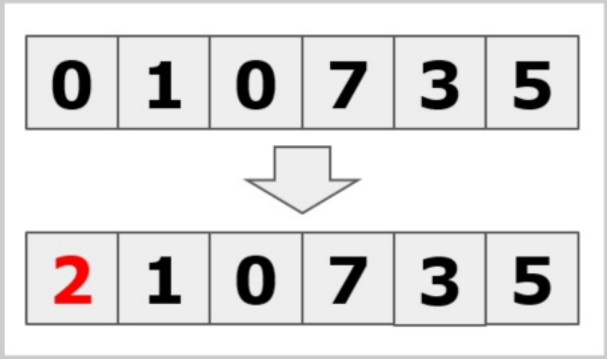
3
攻撃手法

4
まとめ

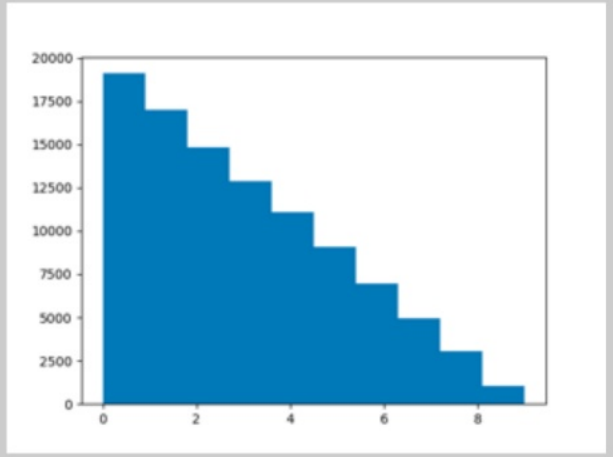
予備戦の加工手法

公開データのみを**匿名化**(=異なる属性値に置換)

- 公開データを完全に匿名化できれば、秘密データの推定は不可能
- 公開データの**属性値分布**をできるだけ保持



匿名化の例

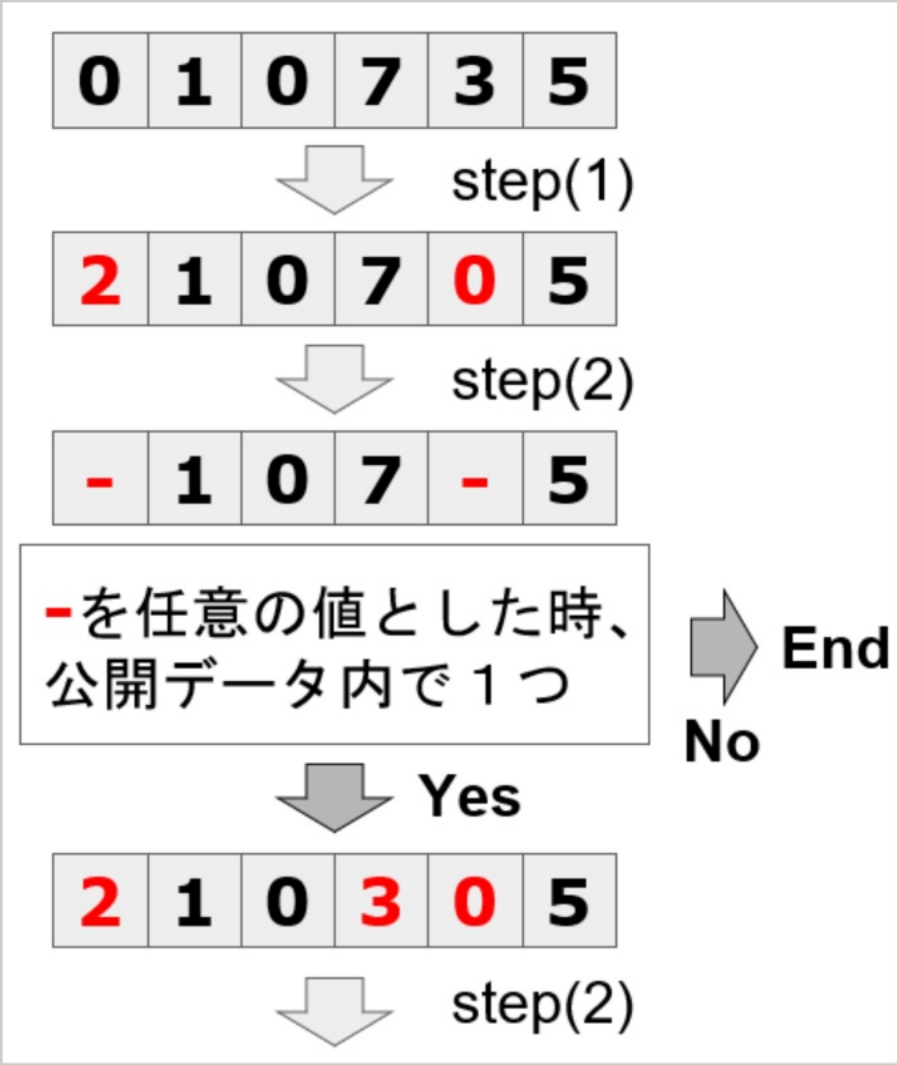


公開データの属性値分布

加工の流れ

(1) **２属性ランダム置換**
　→ 先頭属性とランダムな１属性を匿名化
(2) **一意なレコード**は一意でなくなるまで
　さらに**ランダム置換**
　→ 最大４属性をランダム置換

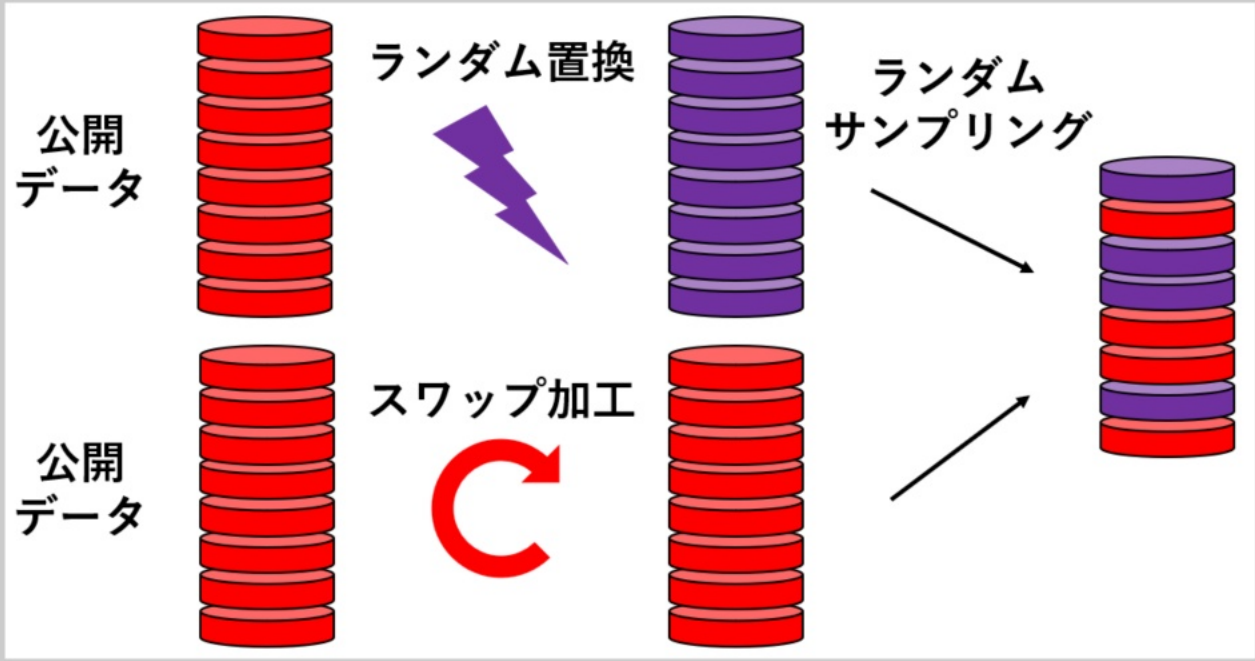
有用性：0.8742
加工部門で１位！



本戦の加工手法

予備戦の他チームの結果から、**スワップ加工**も有効であると分かった

- 本戦は、**(A)ランダム置換**と**(B)スワップ加工**で匿名化したデータを１：１でランダムサンプリング
- **有用性：0.8456**



(A)ランダム置換の流れ

(1) **１属性ランダム置換**
　→ 先頭属性以外のランダムな１属性を匿名化
(2) **一意なレコード**はさらに**ランダム置換**を行う
　→ 最大３属性をランダム置換

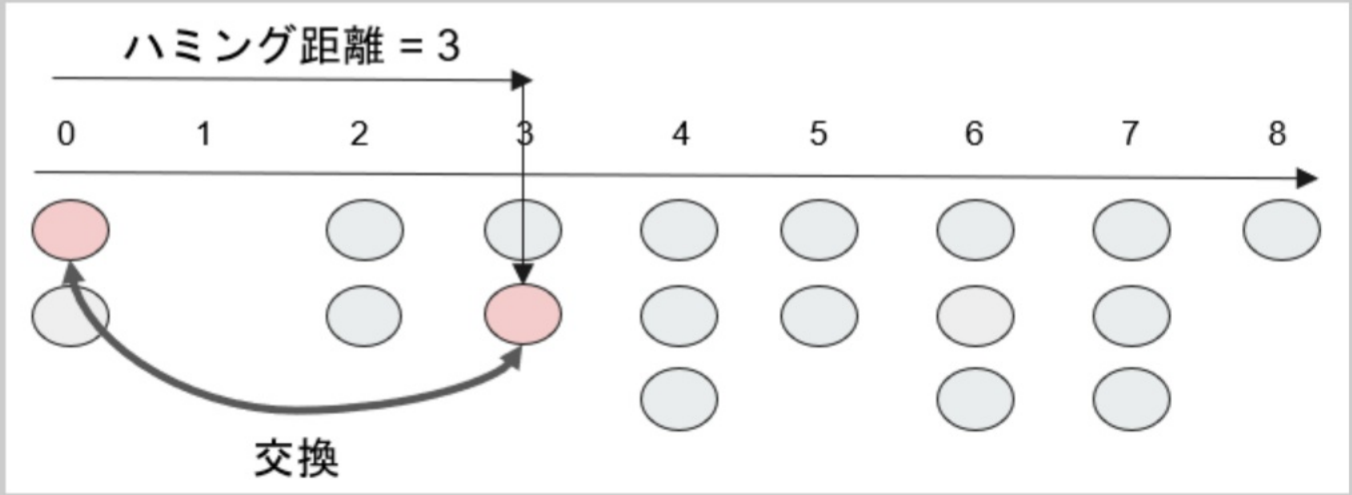
予備戦との違い：(1)で先頭属性を匿名化しない

有用性：0.8567

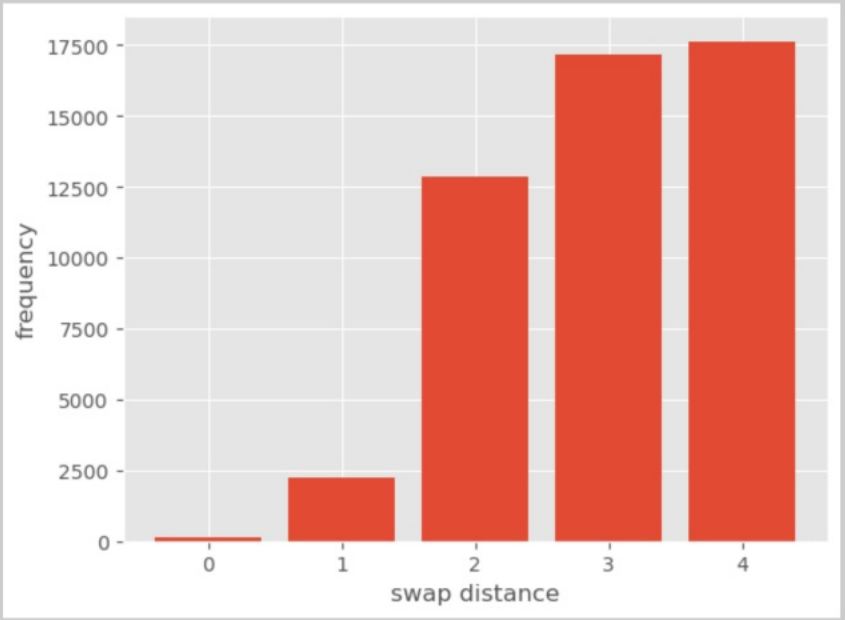
(B)スワップ加工の流れ

ランダムな**ハミング距離**(１～４)で行同士を交換する

有用性：0.8333



ハミング距離３のスワップ



交換したハミング距離の内訳
（０は無交換）

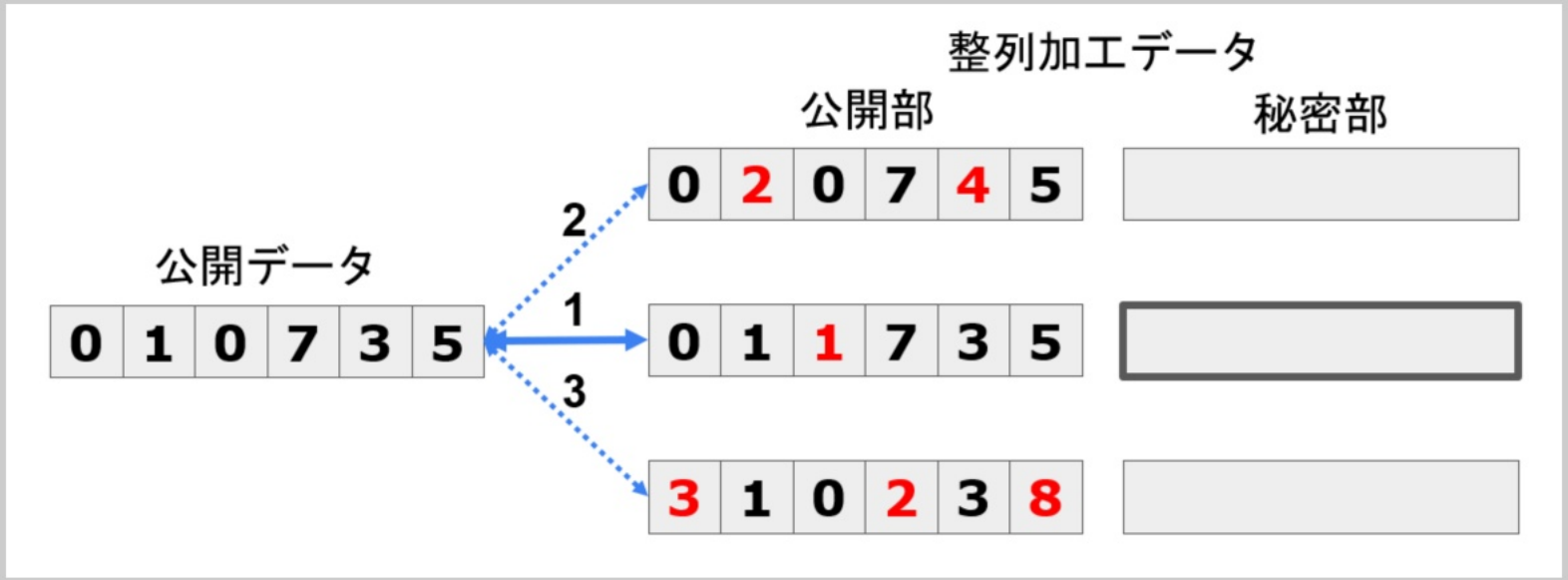
攻撃手法

ベースとなる攻撃手法：**ハミング距離マッチング**

(1) **公開データ**と**整列加工データの公開部**を比較

(2) **ハミング距離最短**のレコード同士をマッチング（重複あり）

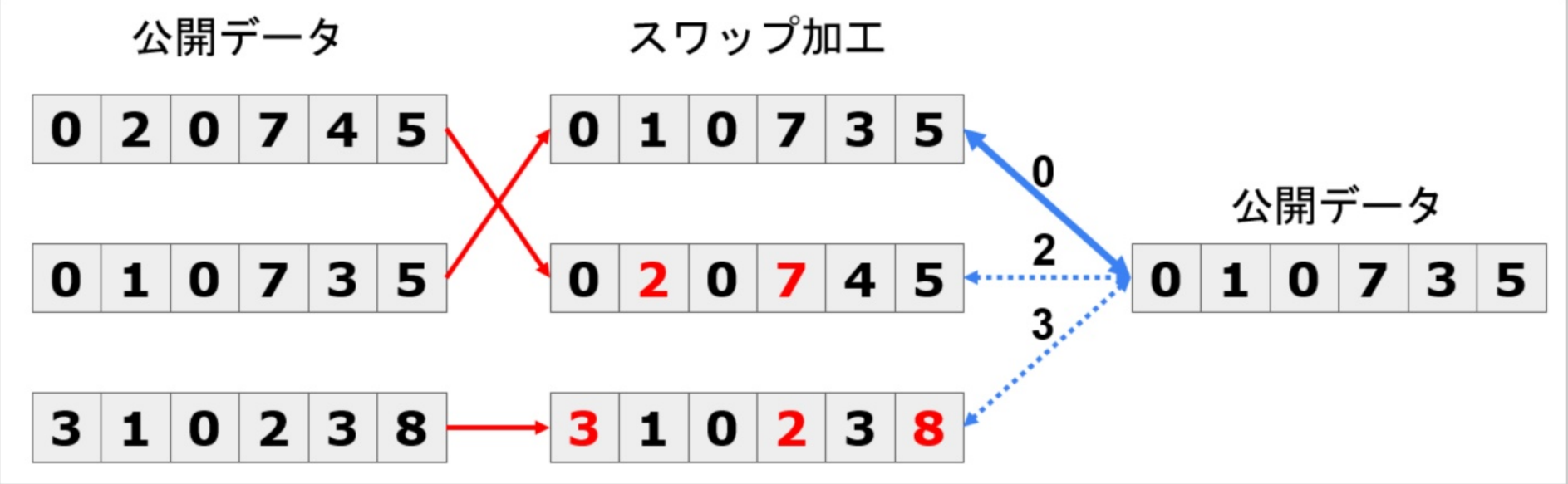
(3) 対応する**整列加工データの秘密部**を提出



攻撃の工夫

ハミング距離マッチングの欠点：スワップ加工したデータに対応できない

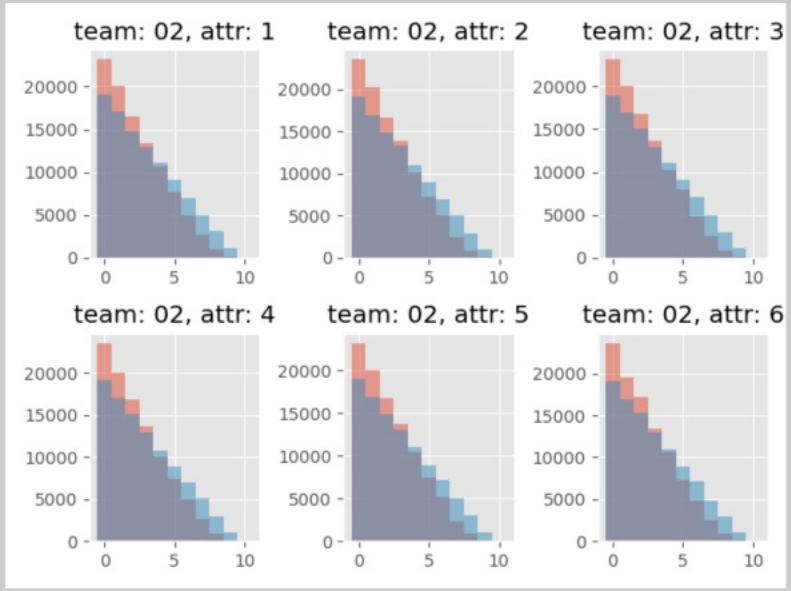
- マッチングを行うハミング距離の**最小値**の選択が重要
- 整列加工データを分析し、チームごとに最適な**最小値**を決定



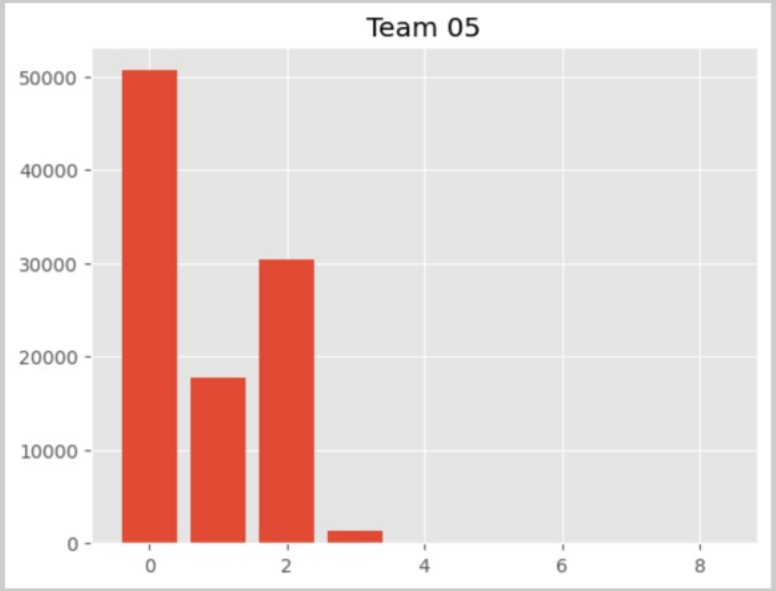
加工データの分析

整列加工データの分析手法（一部）

- 属性値の分布
- 最短ハミング距離の分布　など



(例) 属性値の分布



(例) 最短ハミング距離の分布

まとめ

加工：**ランダム置換**と**スワップ加工**のランダムサンプリング
攻撃：**ハミング距離マッチング**をベースに最適化

感想

- メンバー５人中４人がPWSCup初参加だったが、なかなか健闘できた(と思う)
- 匿名化は奥が深い...

ソースコード