身長の扱いは慎重に！？

慎重な匿名化手法の提案

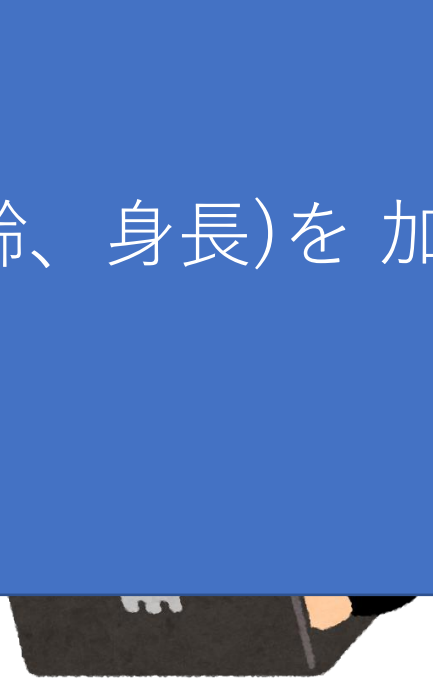
匿名化とは

匿名化とは..

個人が特定される準識別列(今回でいう性別、年齢、身長)を 加工し、
個人を特定できないように加工すること！！

男	20	170.3	×△×

加工後データ



悪用大好きくん

身長がバレるリスク

・身長がどれだけ一意か
・同じ身長のレコードの平均数

レコード数:203521		丸め込み	
		なし	あり
一意のレコード数	身長のみ	16	0
	性年代既知	6247	424
平均レコード数	身長のみ	386.1	3700.4
	性年代既知	6.4	47.5

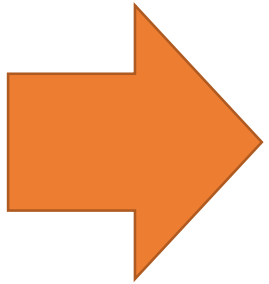
・16人の人は身長のみで特定可能
・性年代を知られていたら6247人も..
・平均でも203521人中6.4人まで絞れてしまう..
→1/6.4の確率で特定可能
・四捨五入しても424人は特定可能
→四捨五入だけでは不十分

提案匿名化手法(夏前ver)

	性別	年齢	身長	病気記録
Aくん	男	20	170.0	○○○
Bくん	男	21	170.1	○×△
Cくん	男	20	170.1	○○△
Dくん	男	22	170.2	○○○
Eくん	男	21	170.3	○△×
Fくん	男	20	170.3	×△×

身長平均化

身長平均化



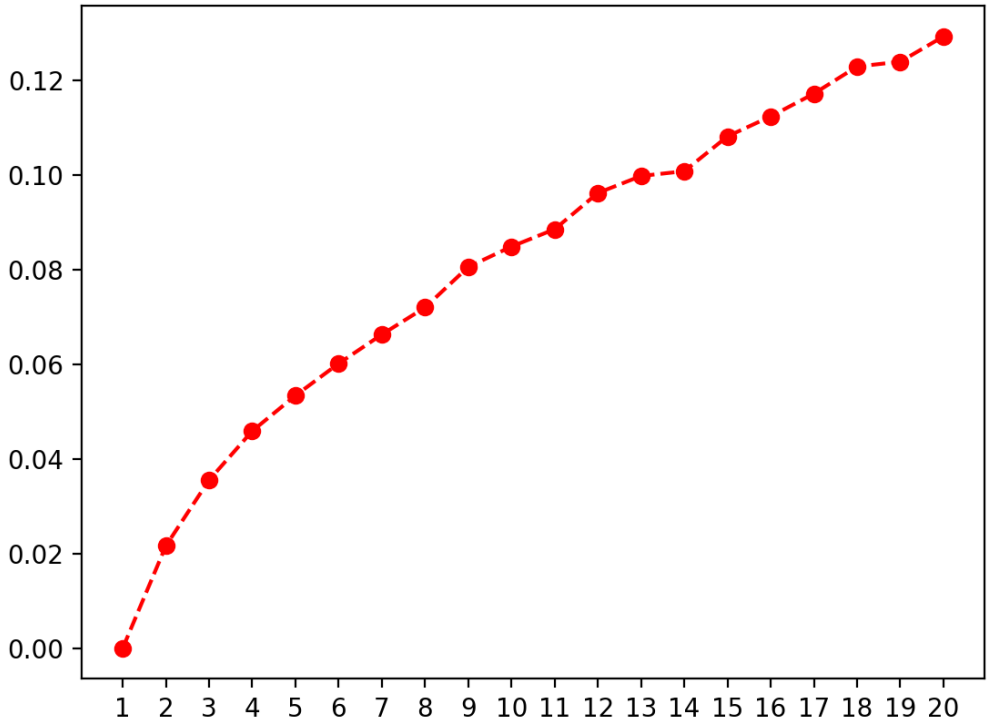
K=3の例

	性別	年齢	身長	病気記録
Aくん	男	20	170.1	○○○
Bくん	男	20	170.1	○×△
Cくん	男	20	170.1	○○△
Dくん	男	20	170.3	○○○
Eくん	男	20	170.3	○△×
Fくん	男	20	170.3	×△×

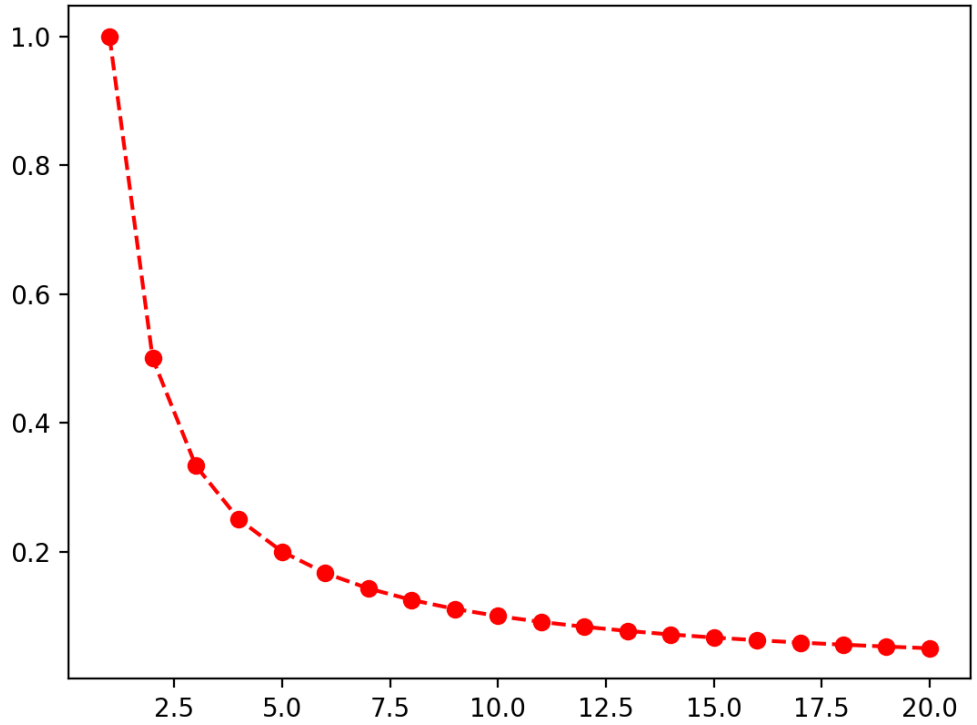
夏はK=3~20まで実装

Kの変化による平均二乗偏差の推移

平均二乗偏差…偏差の二乗の平均の平方根



平均二乗偏差の推移



最高特定率(1/k)の推移

本手法の問題点:身長を全てのデータで加工してしまっている
→すでに一意性の低いレコードは変更する必要ない

提案匿名化手法(夏休み改変ver)

例:k=3の時

	性別	年齢	身長	病気記録
Aくん	男	20	170.0	○○○
Bくん	男	21	170.1	○×△
Cくん	男	20	170.1	○○△
Dくん	男	22	171.2	○○○
Eくん	男	21	172.3	○△×
Fくん	男	20	172.3	×△×

身長四捨五入
年齢３歳ごとに統一

	性別	年齢	身長	病気記録
Aくん	男	20	170	○○○
Bくん	男	20	170	○×△
Cくん	男	20	170	○○△
Dくん	男	20	171	○○○
Eくん	男	20	172	○△×
Fくん	男	20	172	×△×

身長が同じ人が3人以上
いない人はまとめて平均化

	性別	年齢	身長	病気記録
Aくん	男	20	170	○○○
Bくん	男	20	170	○×△
Cくん	男	20	170	○○△
Dくん	男	20	172	○○○
Eくん	男	20	172	○△×
Fくん	男	20	172	×△×

k=3を満たしているためそのまま →

満たしていないためまとめる

身長	人数
170	3
171	1
172	2

身長	人数
170	3
171	0
172	3

匿名化手法結果比較

レコード数:203521		丸め込み	
		なし	あり
一意のレコード数	身長のみ	16	0
	性年代既知	6247	424
平均レコード数	身長のみ	386.1	3700.4
	性年代既知	6.4	47.5

夏前ver

多くのレコードが同じ身長に
変更されてしまっている
→加工しすぎている

下の手法は最小限の加工だが
数値が良い!

レコード数:203521		丸め込み	
		なし	あり
一意のレコード数	身長のみ	0	0
	性年代既知	0	0
平均レコード数	身長のみ	3700.4	3913.3
	性年代既知	153.8	174.4

夏休み改変ver

レコード数:203521		丸め込み	
		なし	あり
一意のレコード数	身長のみ	0	0
	性年代既知	0	0
平均レコード数	身長のみ	425.8	4240.0
	性年代既知	26.5	221.1

条件

丸め込みなしK=10
丸め込みありK=50で加工
身長は3歳ごとに１グループとして加工

夏休み達成事項

〇匿名化全データ実装(K=3~20)
〇実装データ(K=3~20)の平均二乗偏差による有用性評価
〇身長のリスク調査
×身長以外のリスクある属性の調査

課題以外で取り組んだこと
匿名化方法の修正

これから取り組むこと
年齢のグルーピング、kの最適化

80点