K409　菊池研・斉藤研合同発表会　2017年2月4日

# プライバシーを保護した線形回帰システムの実装と評価

明治大学総合数理学部
菊池研4年　濱永 千佳

# 研究背景

- 個人情報保護法改正：要配慮情報（**病歴**、人種、信条など）

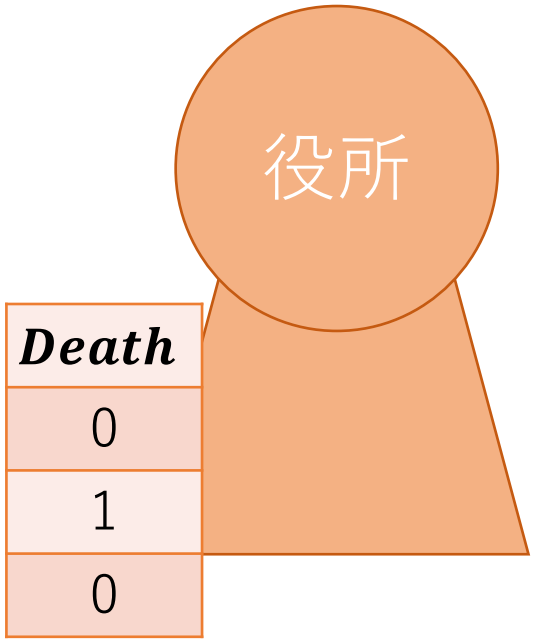
- DPCデータセット（医療データベース）に着目
  - 病気、個人情報、診療情報などからなるデータセット
  - 脳疾患患者のデータ

| 死亡 | 年齢 | 性別 | 意識レベル | がん | 病院の規模 | 肝臓病 | 身体障害 | 脳梗塞の種類 |
|---|---|---|---|---|---|---|---|---|
| 0 | 74 | 0 | 意識有り 1 | 0 | 2 | 0 | 1 | 0 |
| 0 | 55 | 1 | 正常 0 | 0 | 1 | 0 | 4 | 0 |
| 1 | 71 | 0 | 意識なし 3 | 0 | 0 | 0 | 5 | 0 |

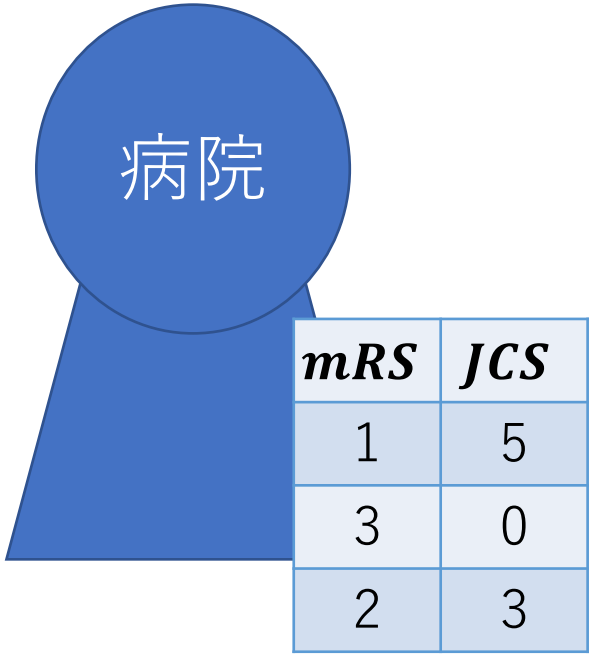
# 問題点：
# データベースが分散

- 死亡診断書など
  →生死の状態

- 病院にいる間のカルテ
  →病気歴や治療の記録

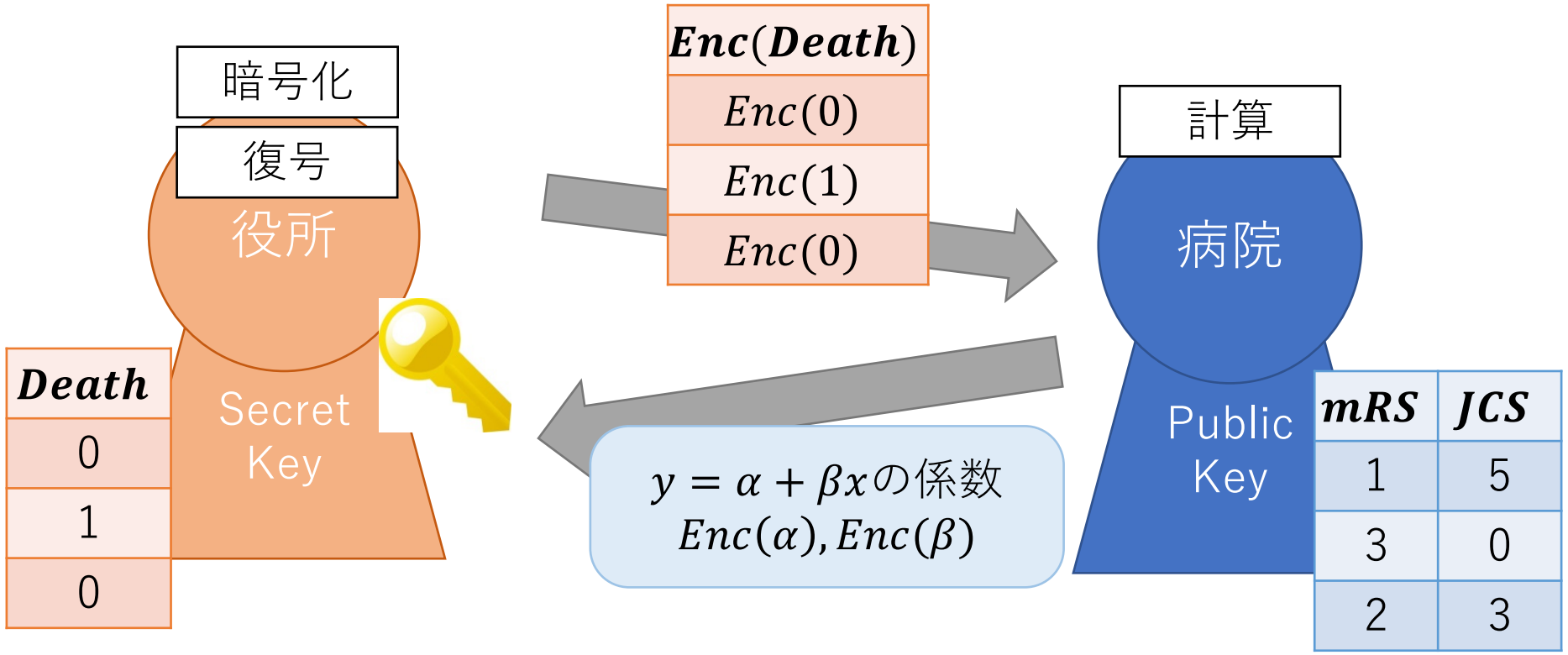



死亡に至った
原因、習慣はなにか？



# 解決方法：
# プライバシー保護データマイニング

- データを暗号化した状態のまま各種の回帰計算を
  行うことにより、データを公開せず、安全に活用する



# 提案手法

- 3種類の秘匿線形回帰プロトコルを提案

| | （1）単回帰 | （2）2変数の重回帰 | （3）多変数の重回帰（n = 2） |
|---|---|---|---|
| モデル | $y = \alpha + \beta x$ | $y = \alpha + \beta_1 x_1 + \beta_2 x_2$ | $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$ |
| BからAへ送るデータ<br>暗号文数 | C,D,E<br>3 | C2,D2,E2, Σ x1, Σ x2<br>5 | F2,G2<br>7 |

# 実験

- 提案手法をscLinearシステムとして実装した.

- DPCデータセットを用いて、scLinearを2点から評価する.
  - 計算の正確性
  - パフォーマンス

表2　実験環境（抜粋）

| | |
|---|---|
| OS | Windows 7 |
| メモリ | 4 GB |
| CPU | Intel® Core™ i5 |
| クロック | 1.8 GHz |
| 使用言語 | Java(1.8.0_91-b14)<br>R(3.1.0) |
| 鍵長 | 2048[bit] |

# 結果：計算結果の正確性

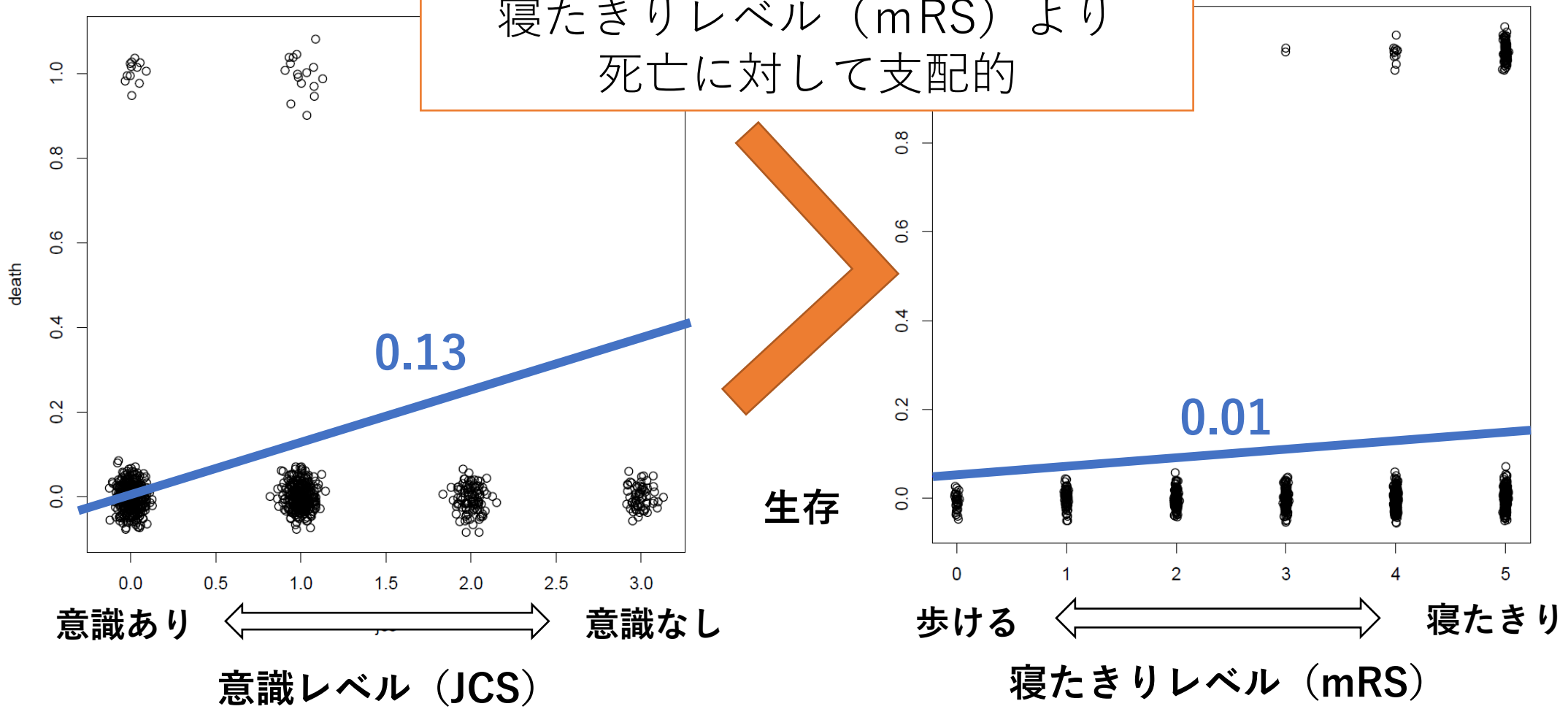
表 3　線形回帰モデルの係数と提案方式の比較 ($n = 5000$)

| | 提案方式 | | R | | | |
|---|---|---|---|---|---|---|
| variables | scLinear | coefficient | Std. Error | t value | $Pr(>\|t\|)$ | |
| $\alpha$ | −0.1731982 | −0.1731982 | 0.0290099 | −5.970 | $2.53e-09$ | *** |
| Age | 0.0015410 | 0.0015410 | 0.0003576 | 4.310 | $1.67e-05$ | *** |
| Sex | −0.0217865 | −0.0217865 | 0.0083993 | −2.594 | 0.009519 | ** |
| JapanComaScale | 0.1283596 | 0.1283596 | 0.0049296 | 26.039 | $< 2e-16$ | *** |
| modifiedRankinScale | 0.0121227 | 0.0121227 | 0.0034845 | 3.479 | 0.000507 | *** |
| StrokeType | 0.0292522 | 0.0292522 | 0.0073582 | 3.975 | $7.12e-05$ | *** |
| LiverDisease | 0.0095770 | 0.0095770 | 0.0324591 | 0.295 | 0.767970 | |

- 提案方式と統計ソフトRの結果に、差は見られなかった
  （*n=1000, 2000の場合においても差がなかった*）
  →提案システムは正確に計算できている

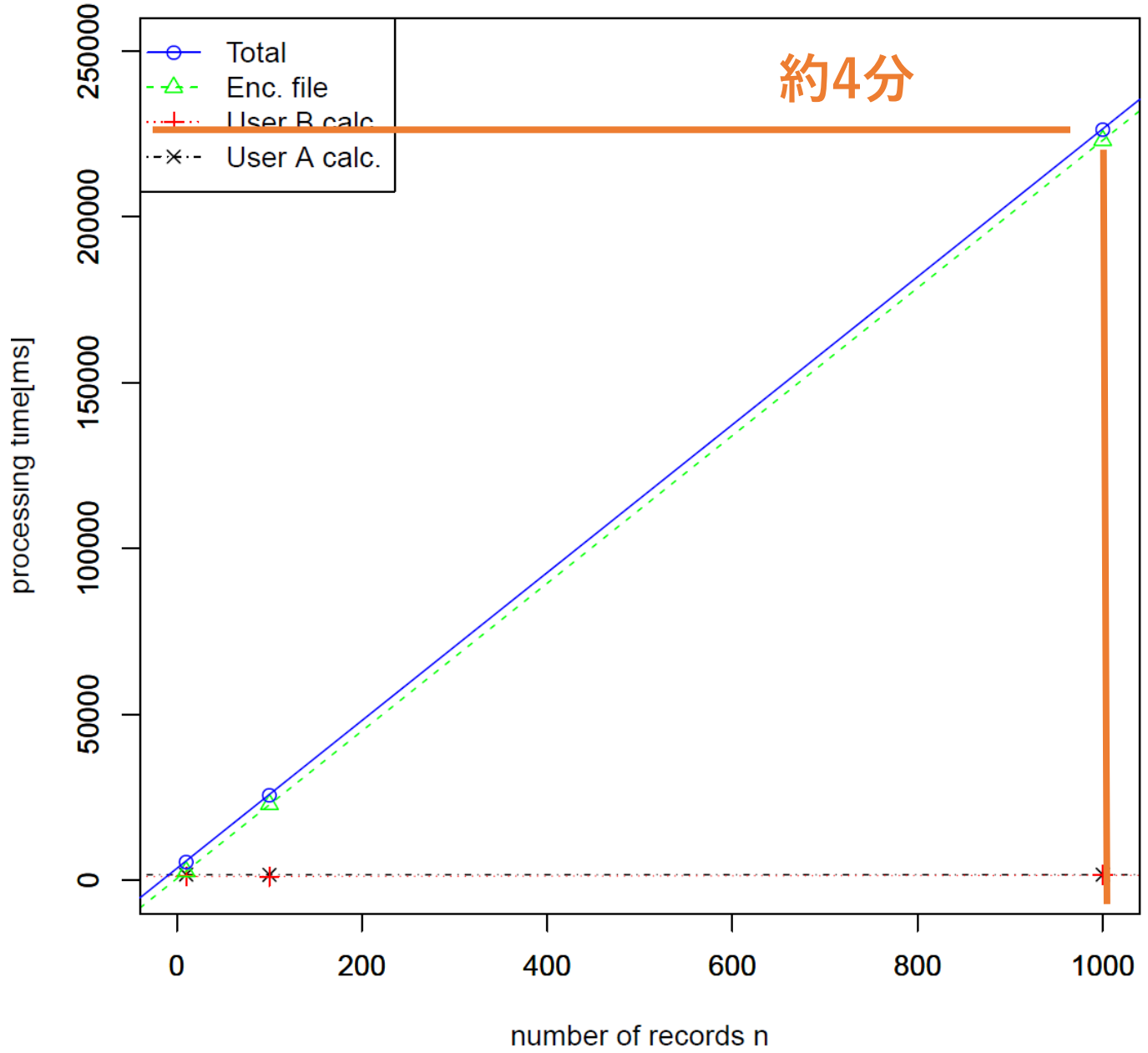
# 実験2　結果：線形回帰直線

意識レベル（JCS）が
寝たきりレベル（mRS）より
死亡に対して支配的



# パフォーマンス

- scLinearでのシステム全体の実行時間は約4分



# 改良の検討

- 統計量の公開なく行いたい
- HElib(完全準同型暗号ライブラリ)を
  用いて処理時間を推定

- [推定より]
- 暗号、計算の処理時間がscLinearより長くなる
- 計算処理の高速化が課題

表 5　処理時間の比較 [sec]($m = 6$, $n = 1000$)

| | scLinear | HElib での推定 |
|---|---|---|
| 暗号化 | 223.742 | 749.262 |
| 計算 | 6.010 | 783.765 |
| 復号 | 5.736 | 4.354 |
| 合計 | 235.488 | 1537.382 |

# おわりに

- 2組織間におけるプライバシーを保護した線形回帰を求める3つのプロトコルを提案した
- scLinearシステムの正確性として、小数第6位までの有効精度であることを示した.

- 秘匿情報の安全を守るために、HElib を用いたシステムを実装する場合における, 計算処理の高速化を検討した.