NIIクラウド上の秘密計算システムを用いた教務データの分析

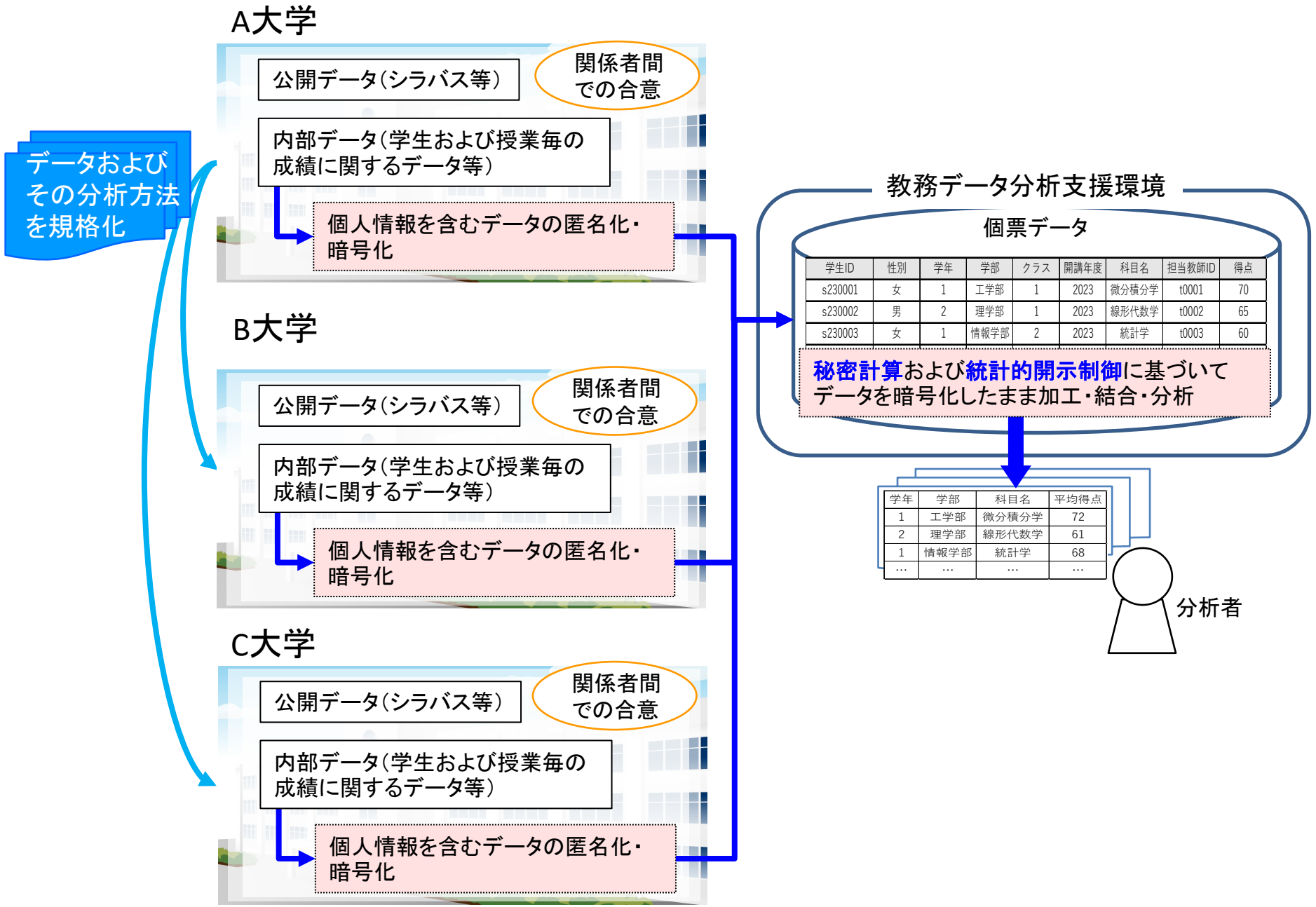
高木理
群馬大学 情報学部
2025年6月18日

教育データ上のプライバシー保護問題

- 教育データを活用することの機運は高まりつつある
  - GIGAスクール構想→教育データ標準
    →教育データ利活用ロードマップ，
    行政機関等匿名加工情報制度
- 一方で，教育現場におけるデータの利活用，特に，研究利用等の二次利用は，期待以上に進展していない

- 教育データの直接のデータ提供者である学生をはじめ，教職員を含む関係者全員のプライバシー保護を考慮する必要がある

秘密計算および統計的開示制御を用いた教務データ分析支援環境



NIIクラウドを用いた秘密計算システムの評価実験

- 本実験の主旨
  - 教務現場における典型的なユースケースシナリオを立て，そのシナリオに沿った分析を，実在の秘密計算システムおよび教務データを用いて実行することを通じて，秘密計算システムに基づく教務データ分析の実用性を評価する.
  - 分析に際して，なるべく同じ条件下で，秘密計算システムを用いた分析と通常の環境における分析を行い，その分析結果および実行時間を比較する.

- 分析対象データ

A) 2014年度～2023年度のある大学のシラバスデータ(総計:53,134件）

➢ 属性の例

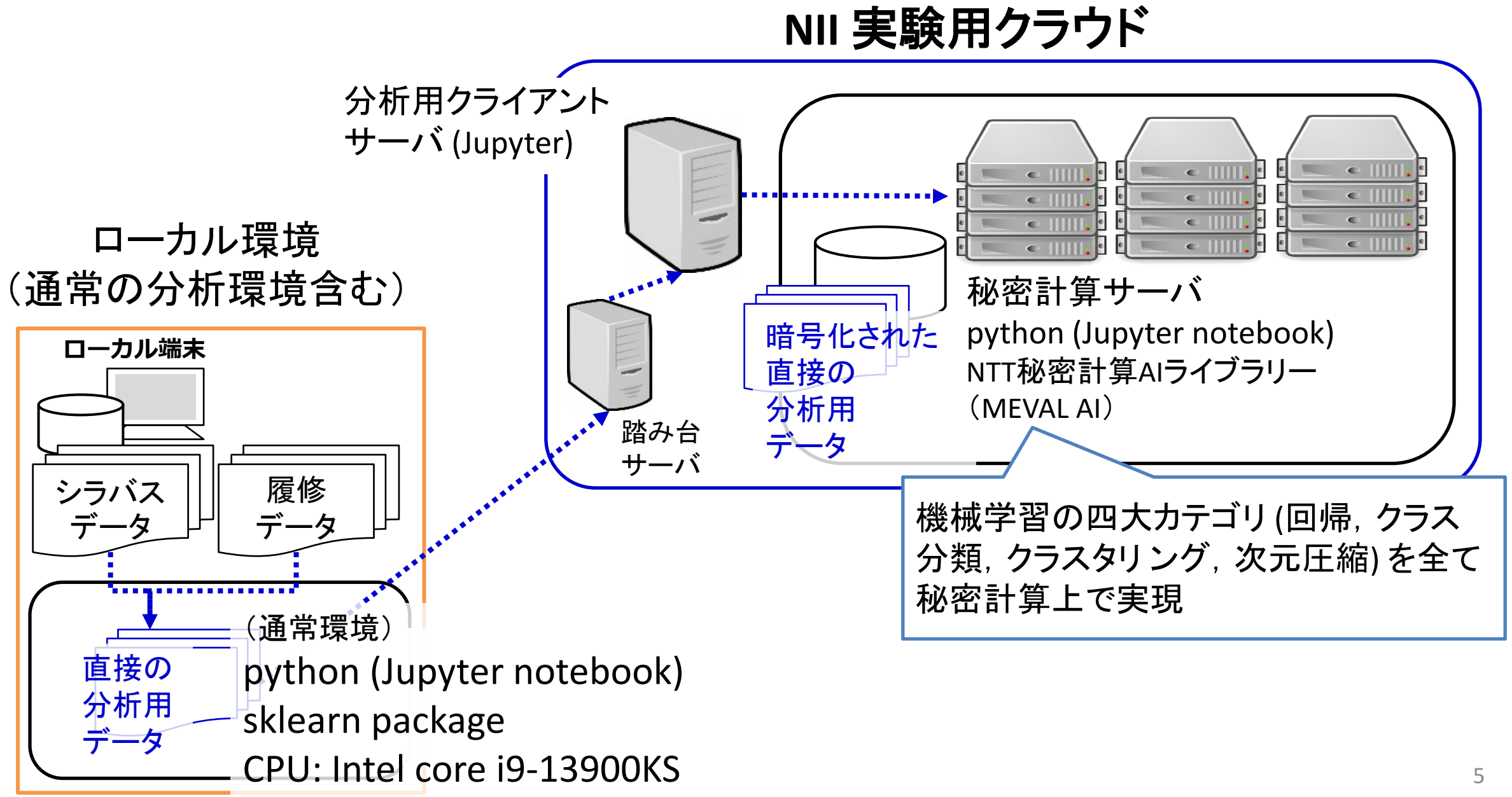
①開講年度，②科目名，③単位，④開講学期名，⑤開講学部名，⑥コース，⑦学部学科名，⑧曜日・時限，⑨対象年次，…

B) 2014年度～2023年度の履修状況に関するデータ（総計:109,8391件）

➢ 属性

①学部等，②年次，③学籍番号（ハッシュ化），④履修年度・学期，⑤時間割コード，⑥授業科目名，⑦授業題目名，⑧単位数，⑨開講年度・曜日・時限

実験環境


NII 実験用クラウド
分析用クライアント
サーバ (Jupyter)
秘密計算サーバ
python (Jupyter notebook)
NTT秘密計算AIライブラリー
（MEVAL AI）
暗号化された
直接の
分析用
データ
踏み台
サーバ
ローカル環境
（通常の分析環境含む）
ローカル端末
シラバス
データ
履修
データ
直接の
分析用
データ
（通常環境）
python (Jupyter notebook)
sklearn package
CPU: Intel core i9-13900KS
機械学習の四大カテゴリ (回帰，クラス分類，クラスタリング，次元圧縮) を全て秘密計算上で実現

ユースケースシナリオ

- 大学学部における1～3年次の履修状況が，4年次以降の履修状況にどのくらい影響を与えるのかを分析したい.

- そこで，以下の3種類の機会学習アルゴリズムを用いて，対象データの各学生の履修状況に関する分類および回帰を行う.
  ① 教師あり分類アルゴリズム（ロジスティック回帰，決定木，GBDT）
  ② 教師あり回帰アルゴリズム（決定木，GBDT，LASSO，FFNN）
  ③ 教師なし分類アルゴリズム（k-means）

① 教師あり分類アルゴリズムによる分析

- 学生毎に，各学年における履修数および再履修数（来年度に再度履修することになる数）を特徴量として，卒業までの在学年数が4年を超えたかどうかを判定する.
- 分析対象
  - ✓ 医学部以外の休学せずに4年以上在籍した卒業生
  - ✓ 本発表では，2つのグループのサイズを揃えた（学生数の補正）
  - ✓ そのため，サイズ（件数）は584になっている.

分析結果の比較(① 教師あり分類アルゴリズム)

ロジスティック回帰（通常環境）				ロジスティック回帰（秘密計算システム）			
	正確性	モデル構築時間	予測時間		正確性	モデル構築時間	予測時間
1年次のみ	49.78%	0.0017	0.0004	1年次のみ	49.43%	2.97	1.41
1～2 年次	77.74%	0.0024	0.0004	1～2 年次	78.00%	5.19	1.41
1～3 年次	80.88%	0.0026	0.0004	1～3 年次	80.23%	5.99	1.39

決定木（通常環境）				決定木（秘密計算システム）			
	正確性	モデル構築時間	予測時間		正確性	モデル構築時間	予測時間
1年次のみ	49.66%	0.0007	0.0004	1年次のみ	51.31%	12.71	6.4
1～2 年次	74.79%	0.0007	0.0004	1～2年次	76.40%	12.61	6.28
1～3 年次	75.59%	0.0008	0.0004	1～3年次	79.54%	12.57	6.34

GBDT（通常環境）				GBDT（秘密計算システム）			
	正確性	モデル構築時間	予測時間		正確性	モデル構築時間	予測時間
1年次のみ	49.81%	0.0385	0.0006	1年次のみ	50.06%	86.33	13.71
1～2年次	76.48%	0.0412	0.0006	1～2年次	76.40%	86.12	13.83
1～3年次	79.81%	0.0429	0.0007	1～3年次	79.77%	86.56	13.73

- 時間単位：秒
- 繰り返し回数：通常環境=1000回，秘密計算システム=10回

② 教師あり回帰アルゴリズムによる分析

- 1～3年次の履修数および再履修数に基づいて，4年時以降の履修数を推定する.
- 分析対象
  - ✓ 医学部以外の休学せずに4年以上在籍した卒業生
  - ✓ サイズ（件数）:5590
- 分析結果の抜粋

	LASSO回帰（α=0.1）に基づく4年次以降の履修数の**推定値**と**実測値**との**相関係数**
1年次の履修数と再履修数に基づくLASSO回帰	0.435
1～2年次の履修数と再履修数に基づくLASSO回帰	0.695
1～3年次の履修数と再履修数に基づくLASSO回帰	0.898

通常環境，繰り返し回数：1000回

分析結果の比較（② 教師あり回帰アルゴリズム）

決定木を用いた回帰（通常環境）			決定木を用いた回帰（秘密計算システム）		
決定係数	モデル構築時間	予測時間	決定係数	モデル構築時間	予測時間
71.33%	0.0028	0.0008	73.84%	13.6068	6.3779

GBDTを用いた回帰（通常環境）			GBDTを用いた回帰（秘密計算システム）		
決定係数	モデル構築時間	予測時間	決定係数	モデル構築時間	予測時間
77.24%	0.1120	0.0019	71.15%	29.3506	7.1874

LASSO回帰（α=0.1）（通常環境）			LASSO回帰（α=0.1）（秘密計算システム）		
決定係数	モデル構築時間	予測時間	決定係数	モデル構築時間	予測時間
80.62%	0.0007	0.0001	77.70%	1.8710	1.2129

FFNNによる回帰（通常環境）			FFNNによる回帰（秘密計算システム）		
MSE	モデル構築時間	予測時間	MSE	モデル構築時間	予測時間
23.93	17.7789	0.4285	26.13	751.0313	11.3967

- 時間単位：秒
- 繰り返し回数：（FFNN以外）通常環境=1000回，秘密計算システム=10回
  （FFNN）通常環境=100回，秘密計算システム=4回

③ 教師なし分類アルゴリズム
k-means法による分類結果の比較

- 教師あり分類アルゴリズムで①で用いた特徴量データを用いて，k-means による学生の二分類を試みる.
- ここで，特に①と同様の学生数を補正した上での分類結果を比較する.

- 分析結果の比較（③ 教師なし分類アルゴリズム）

k-means 法による分類 (通常環境)			
中心点座標 1	中心点座標 2	モデル構築時間	予測時間
(75.88, 5.22)	(84.59, 36.58)	0.0011	0.0005
k-means 法による分類（秘密計算システム）			
中心点座標 1	中心点座標 2	モデル構築時間	予測時間
(75.89, 5.24)	(84.6, 36.73)	11.2747	1.2342

- 時間単位：秒
- 繰り返し回数：通常環境=1000回，秘密計算システム=40回

導入実験に関する考察（まとめ）

- 分析結果自体の比較（通常環境における分析結果との差）:
  - 7項目で，1%未満
  - 5項目で，1～4%
  - GBDTによる回帰：6.1%
  - FFNNによる回帰：8.4%

  まったく同じ条件下で比較できなかったことが原因と考える.

- モデル構築時間および予測時間に関しては，1000～10000倍程度の差があるケースが見受けられた.

- ただし，比較的時間がかかるアルゴリズムに関しては，その差が縮まる傾向があり，教育現場への適用は，少なくとも数百～数千件のオーダーであれば，実用が可能であると考える.

NII-RDCを利用した，より安全な，教務データ分析支援環境の構築に向けて

- 「倫倫姫」を始めとするNII学認LMSは，既に多くの大学に利用され，多大なユーザを有している．
- NII学認LMSを始め，今後，このような大学を超えた教育の共通基盤（共通LMS）が普及していくことが期待される．
- その際，共通LMSの効果等を深く分析していくことが重要である．
- そこで，共通LMSの受講者に関するログデータと，各大学が所有する当該受講者に関する教務データを紐づけ，受講生の行動変容等を分析にすることによって，共通LMSの効果などの分析が可能な枠組みの開発を目指す．