筑波大学
University of Tsukuba

新技術説明会
New Technology Presentation Meetings!

複数機関が持つ秘匿データの安全な統合解析技術

筑波大学
システム情報系　情報理工学位プログラム

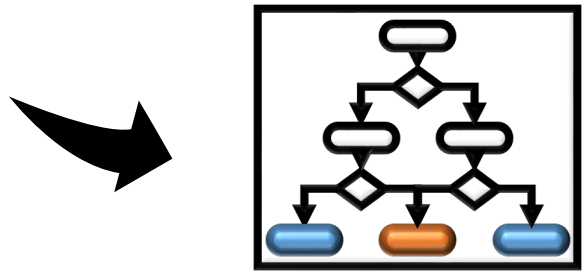
准教授　今倉　暁

2021年10月21日

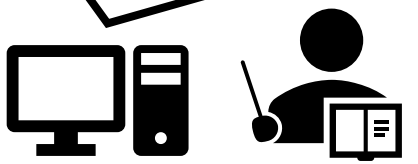
データ解析

- 様々な分野でデータ解析のニーズが高まっている
  - ✓ 医療分野
    - 病気リスク予測
    - リスク因子推定
  - ✓ ものづくり分野
    - 製品開発の最適化
    - 故障検知・予測
  - ✓ 金融分野
    - 需要/リスク予測

ID	Risk	Age	Gender	high	weight	・・・
1	1	45	Male	164.3	65.4	・・・
2	0	25	Female	144.6	46.4	・・・
3	1	36	Female	154.7	43.3	・・・
4	0	62	Male	174.5	73.2	・・・



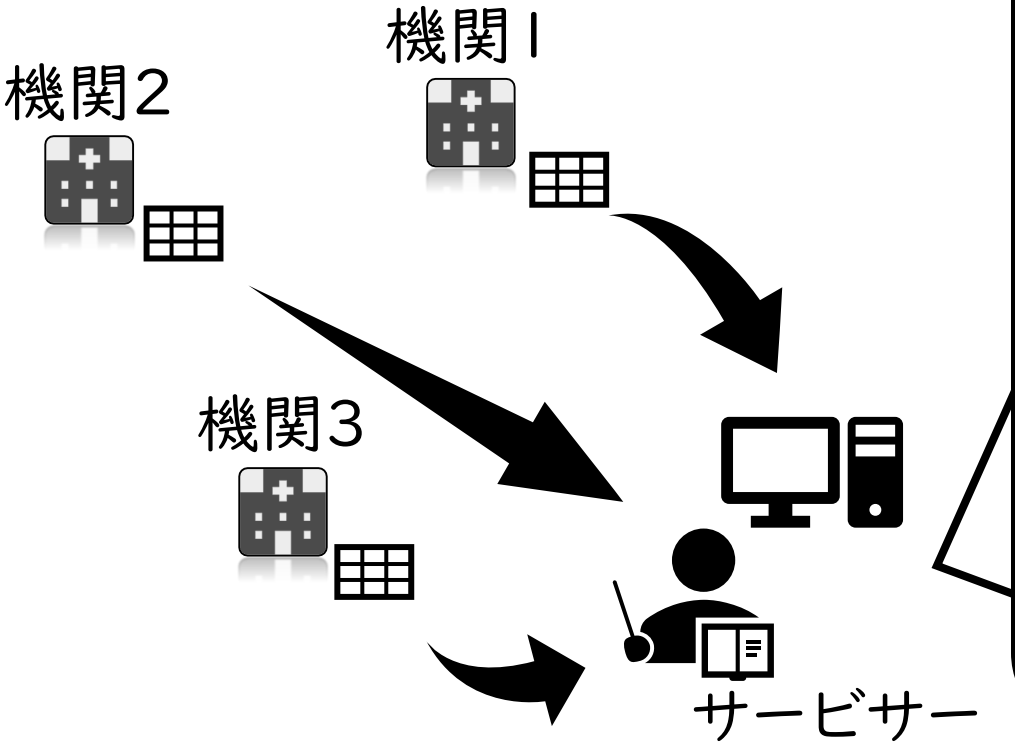
解析モデル



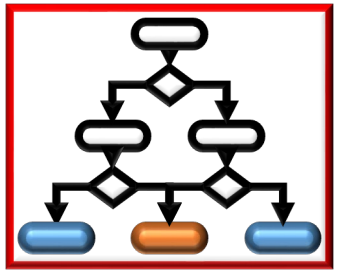
・予測
・因子推定

分散データ統合解析（理想）

- 複数の機関でデータ収集がなされている
  - ✓ これらのデータを統合し解析することで、よりよい解析結果が期待される



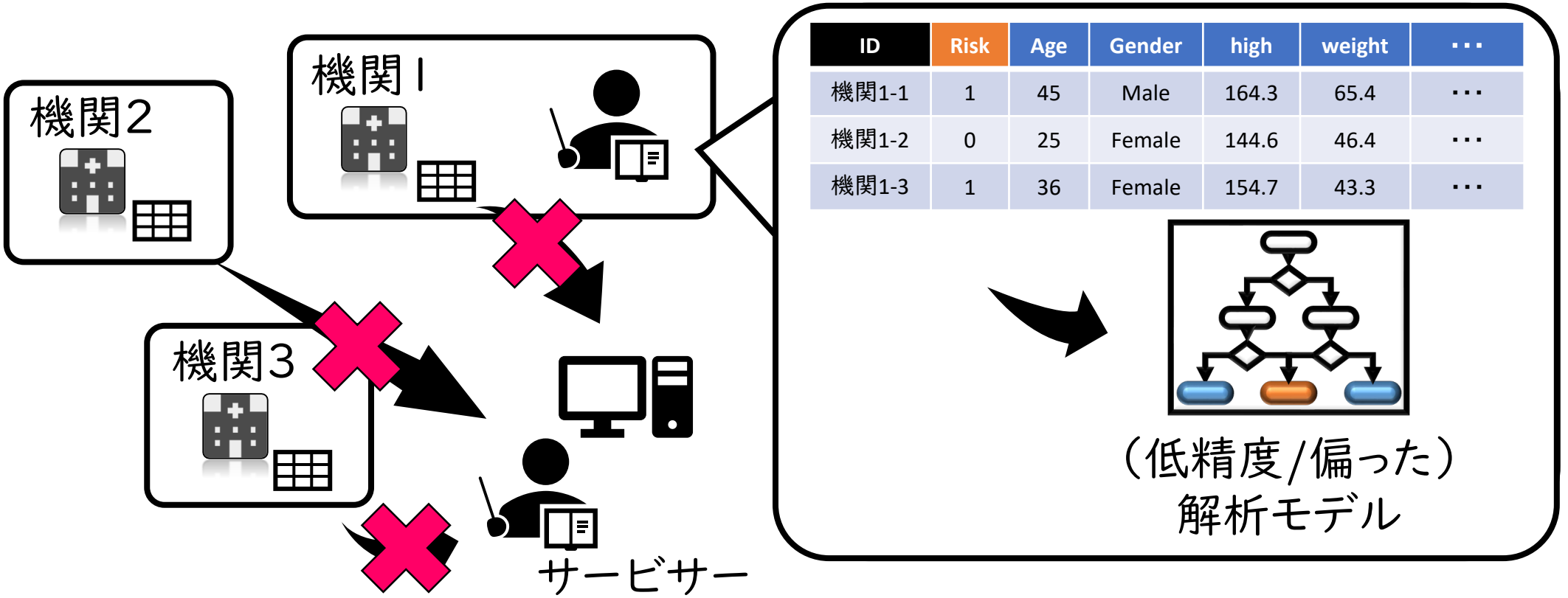
ID	Risk	Age	Gender	high	weight	・・・
機関1-1	1	45	Male	164.3	65.4	・・・
機関1-2	0	25	Female	144.6	46.4	・・・
機関1-3	1	36	Female	154.7	43.3	・・・
機関2-1	0	62	Male	174.5	73.2	・・・
機関2-2	1	12	Male	153.1	76.2	・・・
機関3-1	1	62	Female	174.1	42.5	・・・
機関3-2	0	78	Female	156.8	63.2	・・・



高性能
解析モデル

分散データ統合解析（現実）

- 個人情報保護や企業秘密などの観点から、データを共有することは困難な場合がある
  - → 単独機関のデータのみを用いて解析が行われる

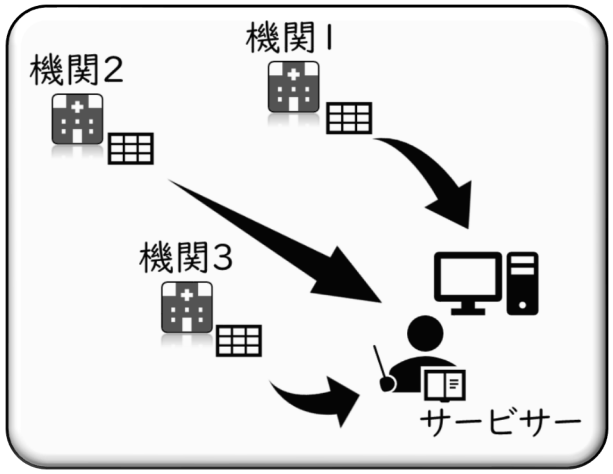


開発技術

データコラボレーション（DC）技術

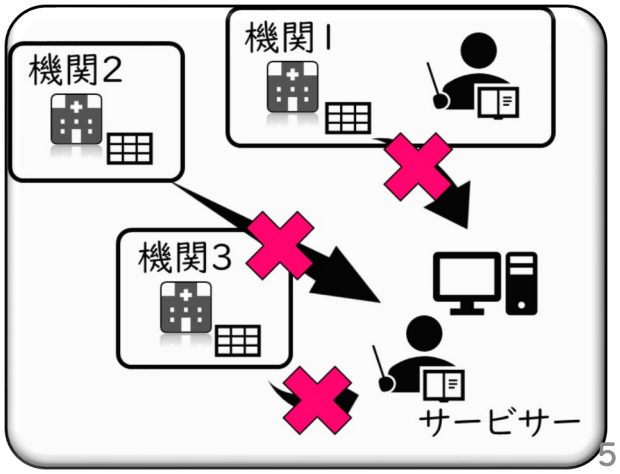
複数機関が持つ秘匿データに対し
「安全に」（秘匿データを他者に公開することなく）
高精度な統合解析を実現する

理想



分散データ統合解析の困難さを克服

現実



スライド構成

- 背景：分散データ統合解析

- データコラボレーション（DC）解析
  - ✓技術概要
  - ✓想定される応用

- 実用化に向けて

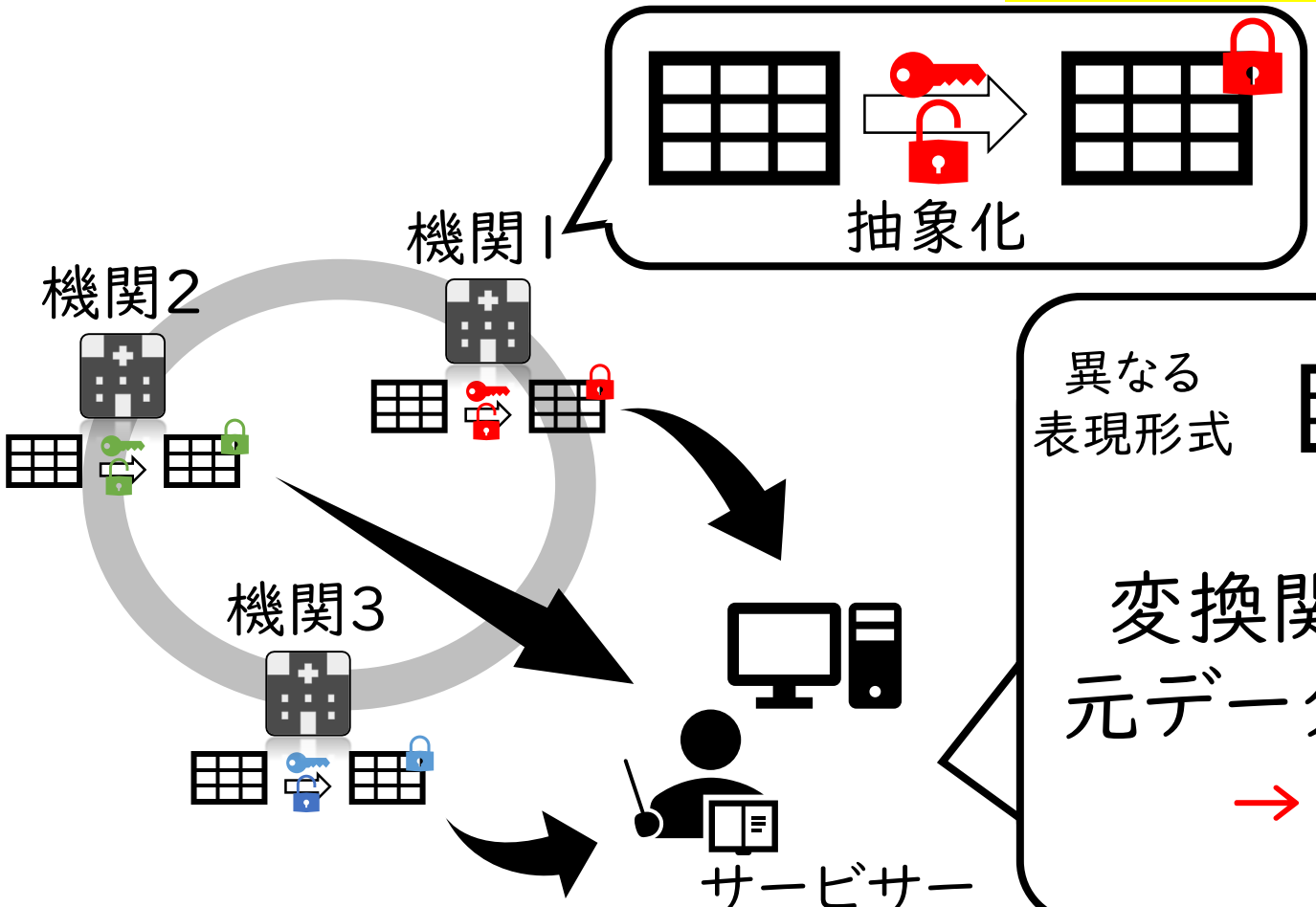
- まとめ

開発技術
データコラボレーション（DC）解析

- 技術概要
- 想定される応用

開発技術
データコラボレーション（DC）技術 [Step I]

- 秘匿情報を含む「生データ」ではなく、各機関が独自に抽象化を行った「中間表現」（🔒）を共有

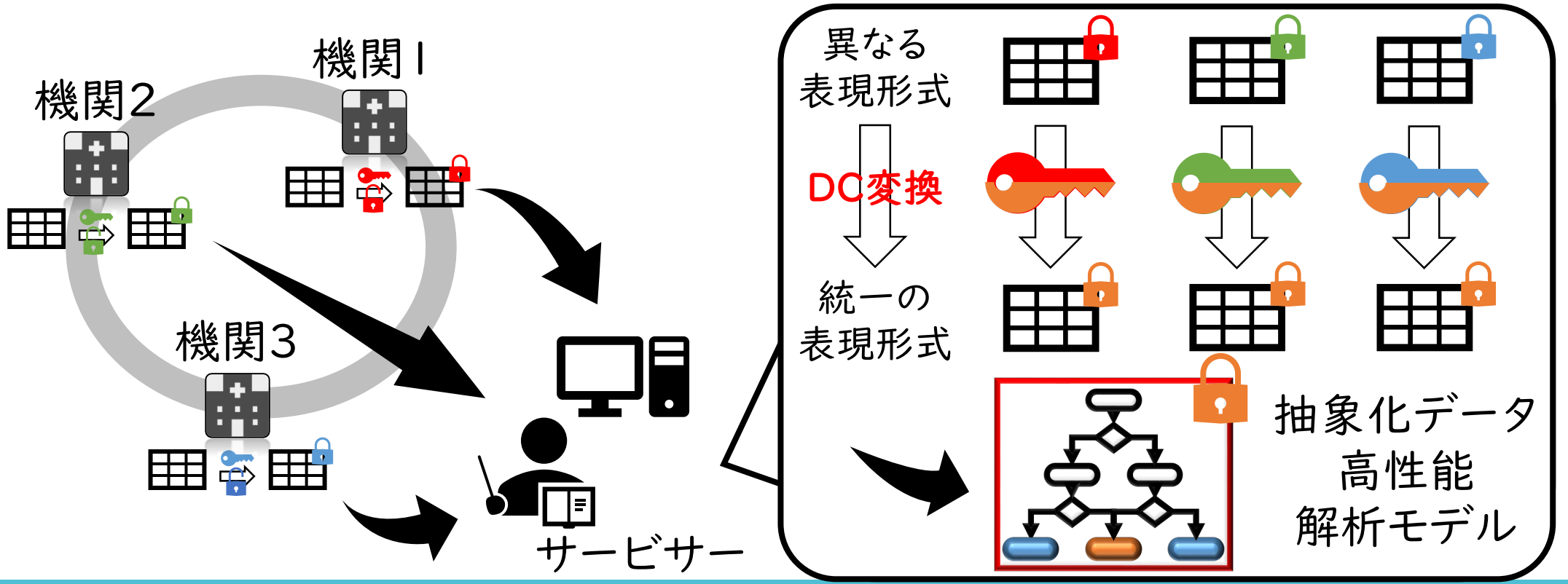


異なる表現形式

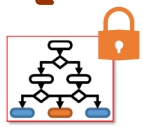
変換関数が不明なため、
元データの復元は出来ない
→ 安全性を担保

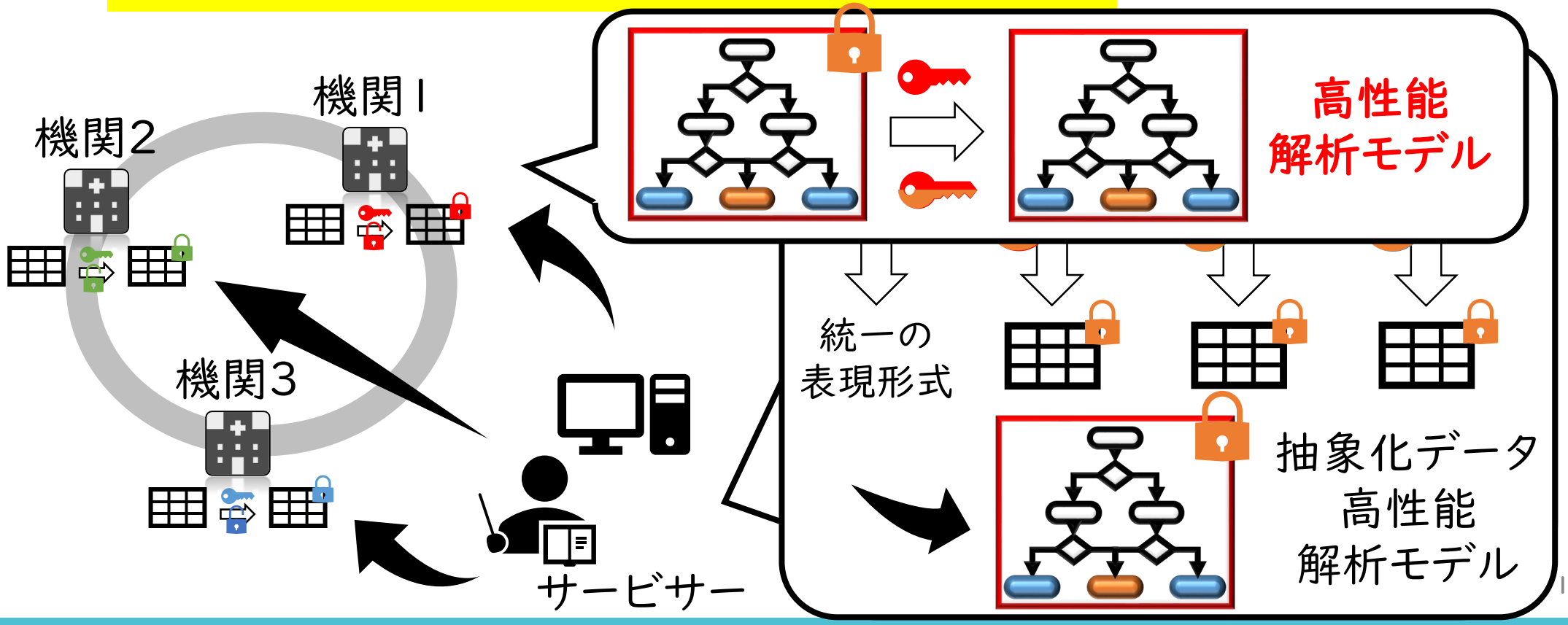
開発技術
データコラボレーション（DC）技術 [Step2]

- 各中間表現（  ）を統合解析可能な統一の表現形式「DC表現」（  ）に変換し、解析
- 抽象化データに対する高性能解析モデルを得る



開発技術
データコラボレーション（DC）技術 [Step3]

- 抽象化データに対する高性能モデル（  ）およびDC変換（  ）を各機関に送付
- 各機関は高性能解析モデルを得る



DC技術でできること

- DC解析ユーザー（各機関）の視点
  - ✓秘匿情報を含む生データを開示することなく、単独では実現出来ないより多くのデータから高性能な解析結果を得ることができる
    - 解析モデル　→ リスク予測
    - 因子推定　→ リスク回避
- DC解析サービス提供者（サービサー）の視点
  - ✓複数機関データを用いた高性能解析の実現
    - → データ提供者を説得しやすい解析サービスの実現
  - ✓新しい価値を生み出すデータ統合パターンの発見
    - → 新しいサービス開発

従来技術とその問題点

- 秘密計算
  - ✓ 暗号化されたデータについて四則演算可能な特殊な暗号化方式（準同型暗号）を用いる
  - → 計算負荷が非常に高く、潤沢な計算資源を用いても大規模データ解析（モデル構築）は困難

- 連合学習（Federated Learning）
  - ✓ 解析モデルを共有し、各機関で順次更新を行う
  - → 機関を跨いだ通信が反復的に必要なため、データの秘匿性が高くネットワークにつなぐことできない場合には適用困難

新技術の優位性

- 準同型暗号を用いないため、計算負荷が小さい
  - ✓各機関の計算負荷は単独機関でのデータ解析と同等
  - ✓サーバーの計算負荷は生データ共有の場合と同等
- 機関を跨いだ通信は解析前後の2回のみである
  - ✓データの機密性が高く外部ネットワークに接続出来ない場合（例、医療データ）でも、人を介したデータ通信により実装可能
- データフォーマットの統一規格化が不要
  - ✓各特徴量の順番や有無、単位系等についての統一規格化をせずに解析することが可能

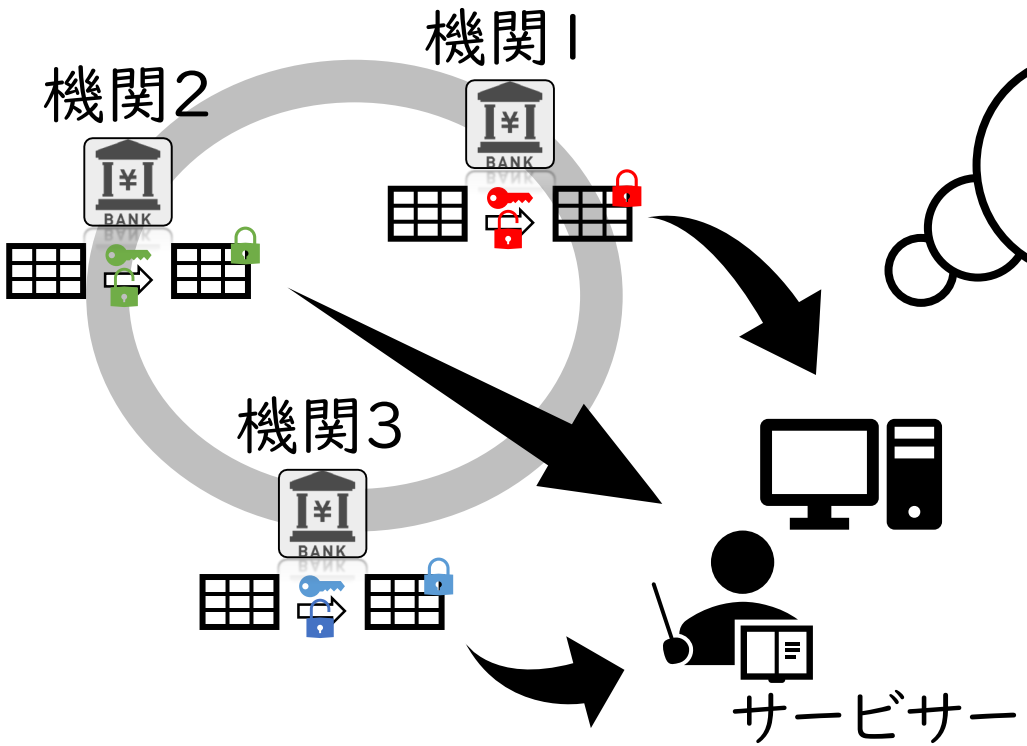
開発技術
データコラボレーション（DC）解析

- 技術概要
- 想定される応用

想定される応用 1/3

- 複数金融機関の持つ顧客企業データの統合解析による資金需要/デフォルトリスク予測

✓複数の金融機関がそれぞれ複数の顧客データを持つ



ID	Risk	資本金	内部留保	時価総額	売上	…
1-1	1	***	***	***	***	***
1-2	0	***	***	***	***	***
1-3	1	***	***	***	***	***
2-1	0	***	***	***	***	***
2-2	1	***	***	***	***	***
3-1	1	***	***	***	***	***
3-2	0	***	***	***	***	***

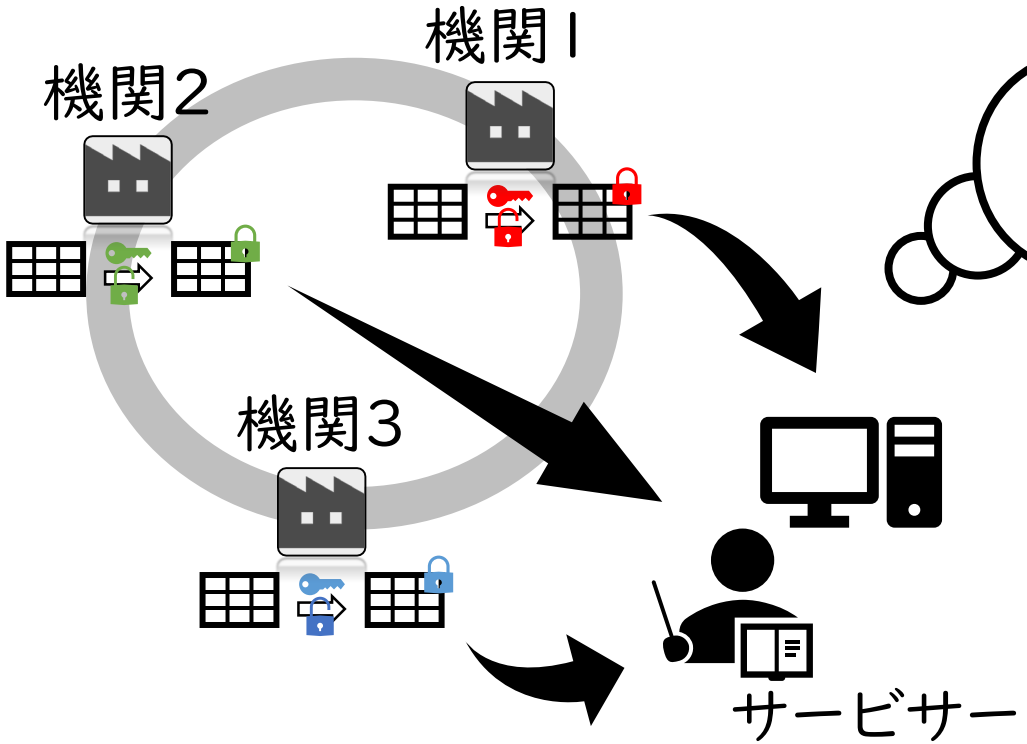
水平分割（サンプルが分散保持されている状況）が想定される

想定される応用 2/3

➢ 製品の製造・運用データ統合解析による故障予測

✓製品データは複数の機関が分散保持する

- 製造時データ：製造企業、運用データ：納入先



ID	Risk	項目1	項目2	項目3	項目4	・・・
1	1	***	***	***	***	***
2	0	***	***	***	***	***
3	1	***	***	***	***	***
4	1	***	***	***	***	***
5	0	***	***	***	***	***

垂直分割（特徴量が分散保持されている状況）が想定される

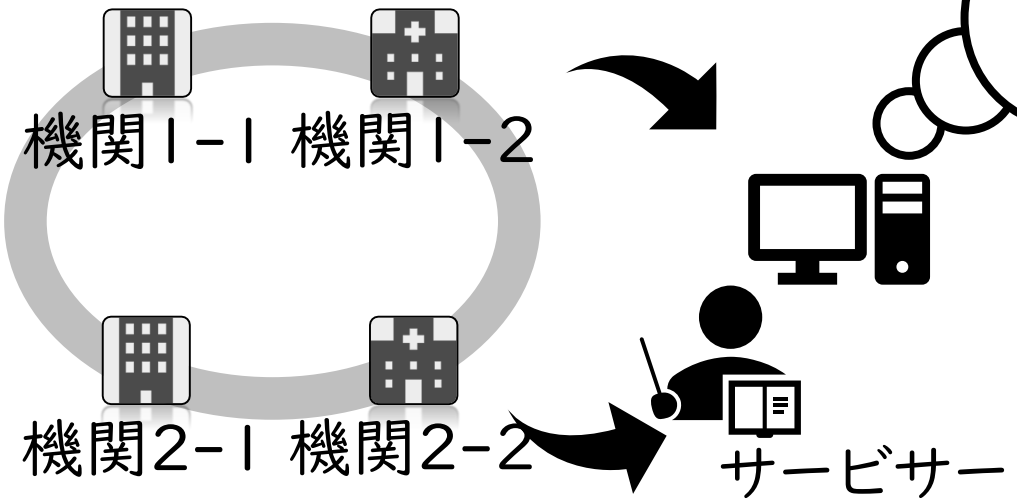
想定される応用 3/3

➢ 複数企業の従業員の医療・健康データの統合解析による健康リスク発見・因子推定

✓複数の企業にそれぞれ複数の従業員がいる

✓各従業員のデータは複数の機関が分散保持する

- 勤怠データ：人事部
- 健康診断結果：病院



ID	Risk	項目1	項目2	項目3	項目4	・・・
1-1	1	***	***	***	***	***
1-2	0	***	***	***	***	***
1-3	1	***	***	***	***	***
2-1	1	***	***	***	***	***
2-2	0	***	***	***	***	***

2次元分割（サンプルおよび特徴量が分散保持されている状況）が想定される

DC技術でできること（再掲）

- DC解析ユーザー（各機関）の視点
  - ✓秘匿情報を含む生データを開示することなく、単独では実現出来ないより多くのデータから高性能な解析結果を得ることができる
    - 解析モデル　→ リスク予測
    - 因子推定　→ リスク回避
- DC解析サービス提供者（サービサー）の視点
  - ✓複数機関データを用いた高性能解析の実現
    - → データ提供者を説得しやすい解析サービスの実現
  - ✓新しい価値を生み出すデータ統合パターンの発見
    - → 新しいサービス開発

適用事例1

➢ 高性能な健康リスク推定モデルの構築

✓腎臓カテーテルデータ

✓特徴量

- 年齢、性別、
- 病気の有無（GN、AN、PKD）、フレイル（虚弱）

✓水平分割（サンプルが分散）

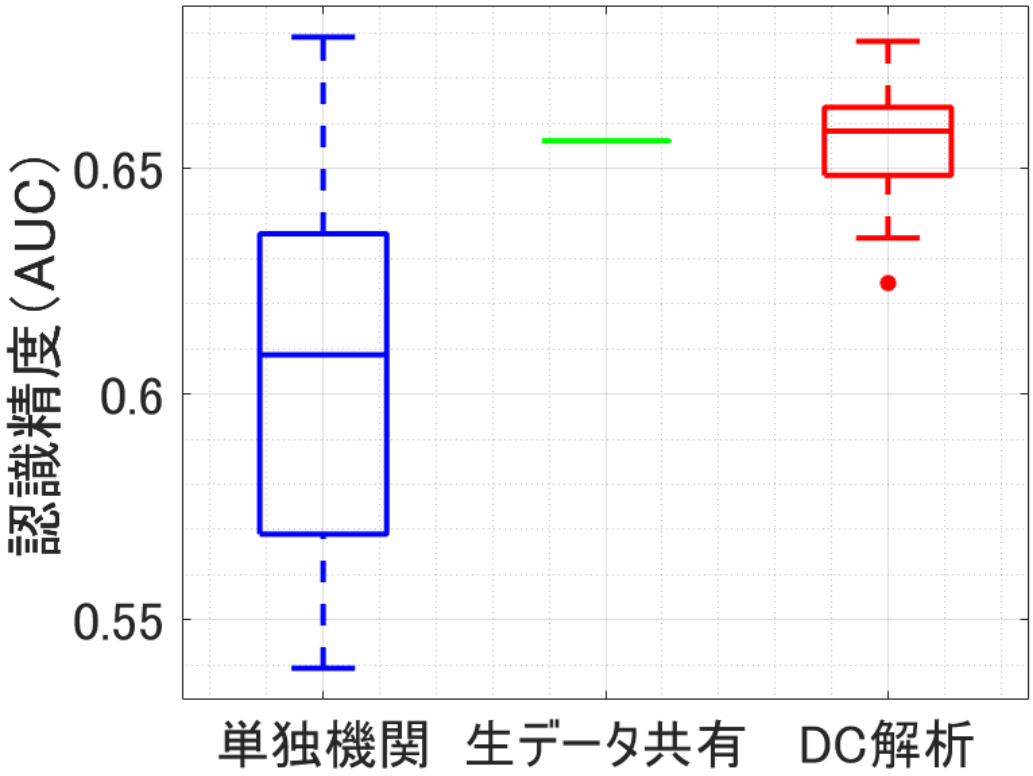
age	sex	GN	AN	PKD	frail	Risk
48	2	1	0	0	1.90	1
44	2	0	0	0	1.30	1
57	2	0	1	0	0.50	0
…	…	…	…	…	…	…
46	1	0	0	1	0.20	1

適用事例1

➢ 高性能な健康リスク推定モデルの構築

✓解析精度

- 単独機関および生データ共有時の精度と比較



生データ共有に近い
認識性能を発揮

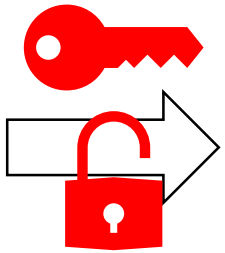
適用事例1

- 高性能な健康リスク推定モデル
  - ✓安全性の確認

中間表現から元データの推測は出来ない

機関1：生データ

age	sex	GN	AN	PKD	frail
48	2	1	0	0	1.90
44	2	0	0	0	1.30
57	2	0	1	0	0.50



機関1：中間表現

F1	F2	F3	F4	F5
-17.60	-1.53	-28.93	-23.76	36.76
-16.23	-0.05	-27.63	-22.88	34.38
-21.25	1.11	-35.79	-29.52	47.85

機関2：生データ

age	sex	GN	AN	PKD	frail
30	2	0	0	0	0.60
43	1	0	1	0	0.70
43	2	0	1	0	1.00

機関2：中間表現

Fa	Fb	Fc	Fd	Fe
-39.85	-18.43	-6.65	-38.56	-43.40
-59.80	-26.86	-11.13	-53.27	-61.30
-58.61	-25.80	-10.99	-54.48	-62.99

適用事例1

➢ 高性能な健康リスク推定モデル

✓安全性の確認

安全性を担保したまま
表現形式を揃える

機関1：中間表現

F1	F2	F3	F4	F5
-17.60	-1.53	-28.93	-23.76	36.76
-16.23	-0.05	-27.63	-22.88	34.38
-21.25	1.11	-35.79	-29.52	47.85

機関1：DC表現

F1	F2	F3	F4	F5
-170.4	-2.49	-0.69	-2.94	0.68
-155.2	0.40	-1.50	0.08	-1.09
-200.8	5.72	-3.72	2.94	0.32

機関2：中間表現

Fa	Fb	Fc	Fd	Fe
-39.85	-18.43	-6.65	-38.56	-43.40
-59.80	-26.86	-11.13	-53.27	-61.30
-58.61	-25.80	-10.99	-54.48	-62.99

機関2：DC表現

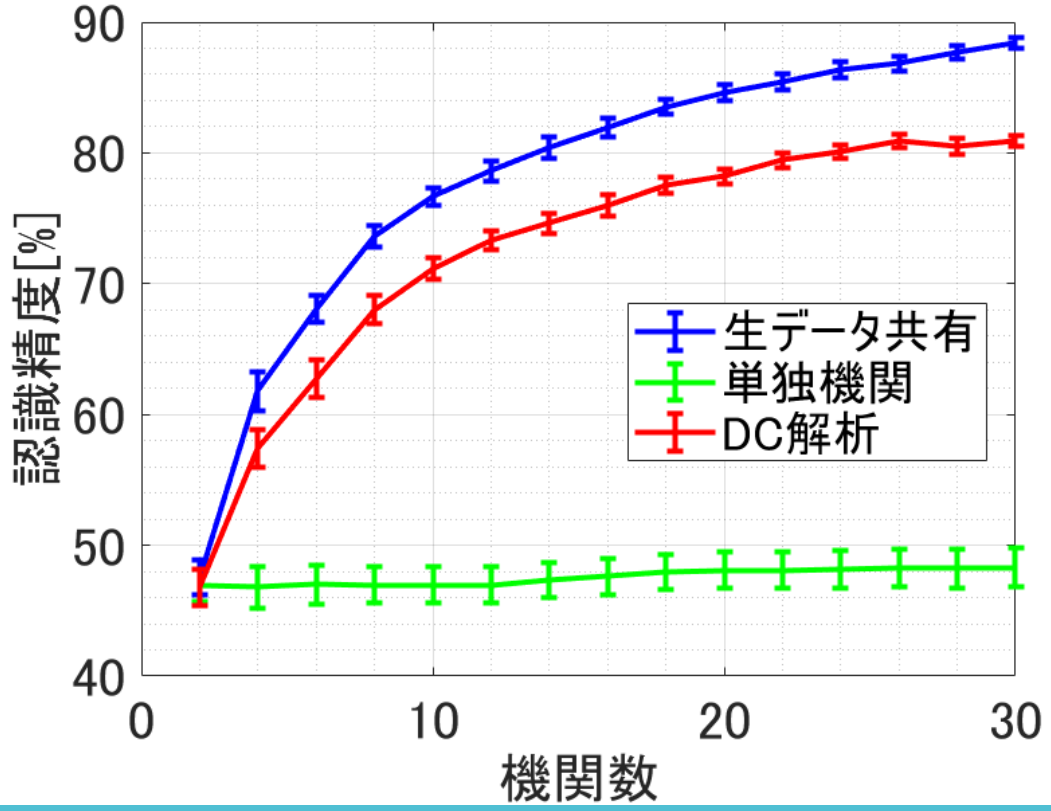
Fa	Fb	Fc	Fd	Fe
-106.6	-0.16	-3.60	-0.47	-1.42
-151.2	4.17	-1.42	3.13	0.97
-152.4	1.29	-3.81	2.87	0.18

適用事例2

➢ 高性能な画像認識モデルの構築

✓手書き数字画像の認識（MNIST）

✓2次元分割（サンプル、特徴量ともに分散）



生データ共有に近い
認識性能を発揮

適用事例3

➢ 企業格付の主要因子の特定

✓企業格付データ（CreditRating_historical.dat）

✓特徴量

- 運転資本/総資産 (WC_TA)
- 内部留保/総資産 (RE_TA)
- 税引前利払前利益/総資産 (EBIT_TA)
- 株式時価総額/全債務の簿価 (MVE_BVTD)
- 売上高/総資産 (S_TA)
- 業種ラベル（Industry）（12種類）

✓目的変数：企業格付（AAA, AA, …, CCC）

適用事例3

➢ 企業格付の主要因子の特定

✓2次元分割（サンプル、特徴量ともに分散）

WC_TA	RE_TA	EBIT_TA	MVE_BVTD	S_TA	Industry	Rating
0.013	0.104	0.036	0.447	0.142	3	{'BB' }
0.232	0.335	0.062	1.969	0.281	8	{'A' }
…	…	…	…	…	…	…
0.096	0.257	0.053	0.968	0.251	11	{'BBB'}
0.204	0.205	0.058	0.83	0.259	2	{'BB' }
0.235	0.227	0.068	0.557	0.148	2	{'BB' }
…	…	…	…	…	…	…
-0.05	-0.023	0.029	0.568	0.139	9	{'BB' }

機関1-1

機関1-2

機関2-1

機関2-2

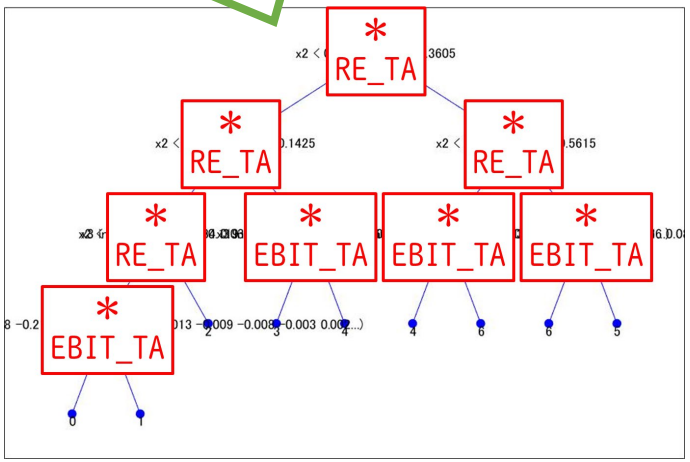
適用事例3

➢ 企業格付の主要因子の特定

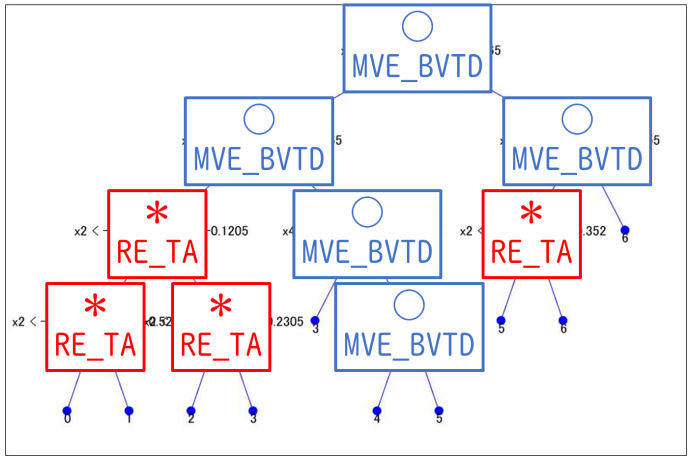
✓各手法で構築された決定木モデル

一部の特徴量のみを持つ単独機関では因子推定に失敗

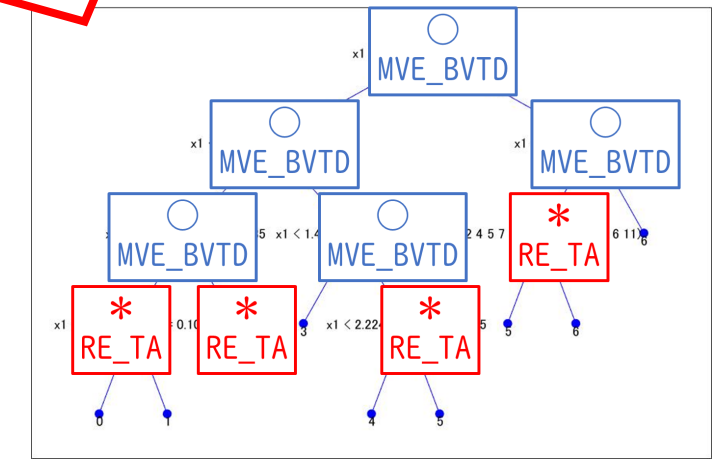
生データ共有の場合と同じ主要因子（RE_TA、MVE_BVTD）の特定に成功



機関1-1での解析



生データ共有



DC解析

実用化に向けて

実用化への課題

- 複数機関の実データでの有効性の実証
  - ✓実データに合わせたデータ前処理技術の開発・選定
  - ✓実ニーズに即した性能評価の実施

- ソフトウェア開発
  - ✓ユーザー向けソフトウェア実装
  - ✓サービス提供者向けソフトウェア実装

企業への期待

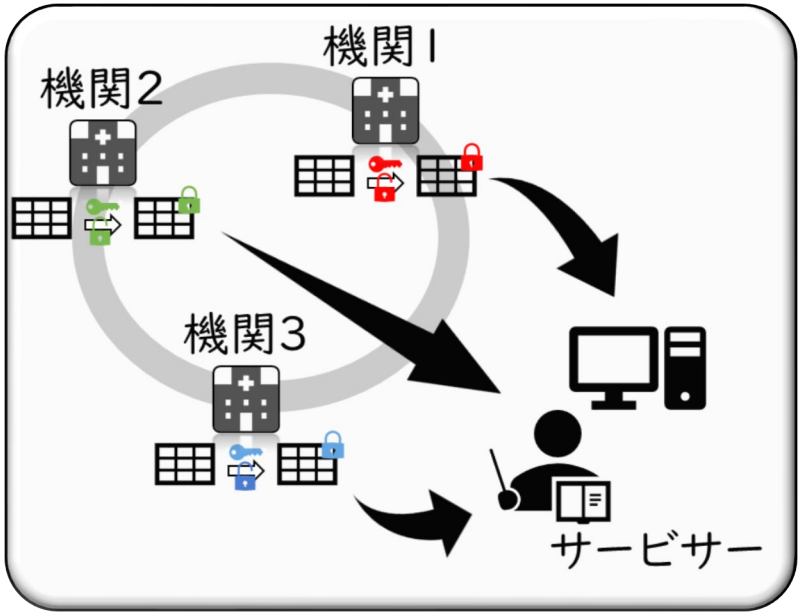
- サービス提供事業者
  - ✓法的な制限や、個人情報・機密保護に敏感な市場におけるサービスニーズの発掘
  - ✓ユーザー企業との連携による実証実験の実施
  - ✓サービス事業の構築

- サービス利用ユーザー企業
  - ✓機会損失を回避すべくサービス事業者との連携
  - ✓実証実験への実データ提供
  - ✓DC解析の事業活用

まとめ

開発技術

データコラボレーション（DC）技術



複数機関が持つ秘匿データに対し「安全に」高精度な統合解析を実現する

分散データ統合解析の困難さを克服

本技術に関する知的財産権

- 発明の名称:
  分散データ統合装置、分散データ統合方法、及びプログラム

- 出願番号　:PCT/JP2019/049551
- 出願人　　:国立大学法人筑波大学
- 発明者　　:今倉暁、櫻井鉄也

お問い合わせ先

筑波大学　国際産学連携本部
クリエイティブマネージャー　國土　順一

TEL　029-859-1490
FAX　029-859-1693
e-mail　kokudo.junichi.ge@un.tsukuba.ac.jp