仮想のパーソナルデータや臨床データを生成するAIとその応用

日本大学　理工学部　応用情報工学科
助教　関　弘翔

2024年12月24日

内閣府｜AI戦略2022

データ関連基盤整備

- AI技術の発展を支えるのは大量のデータである
- データをサイバー攻撃などのリスクから守りながら、分析・解析に活用することは極めて重要である
- 健康・医療・介護、農業等の分野における、AIの活用のためのデータ連携基盤の本格稼働を目標としている

新技術説明会
New Technology Presentation Meetings!

データ連携基盤整備の課題

❶プライバシー保護の必要性
❷十分なデータ量が得られない

➡潜在的リスクにより匿名化だけではデータ共有が進まない
➡そもそも収集できていないデータは共有もできない

新技術説明会
New Technology Presentation Meetings!

データ連携基盤整備の課題

❶プライバシー保護の必要性

- パーソナルなデータは匿名化による保護が一般的である
  ➡ どれほどのプライバシー保護効果があるのか？

- **Netflixの事例**
  - Netflixは1999～2005年の間に約48万人の加入者が映画をレイティングした約1億件のデータを匿名化のうえ、公開した
  - **公開データから個人を特定できる可能性**が指摘された*
    - ある個人が過去に与えた2つのレイティング値およびその日付(誤差3日以内)があれば68%の確率で特定可能である

ユーザID	映画名	評価	評価日
U0001	AA	5	20xx/yy/aa
U0001	BB	4	20xx/yy/bb
U0002	AA	2	20xx/yy/aa

山田太郎さんは
20xx/yy/aaにAA
という映画を見て
5と評価したらしい

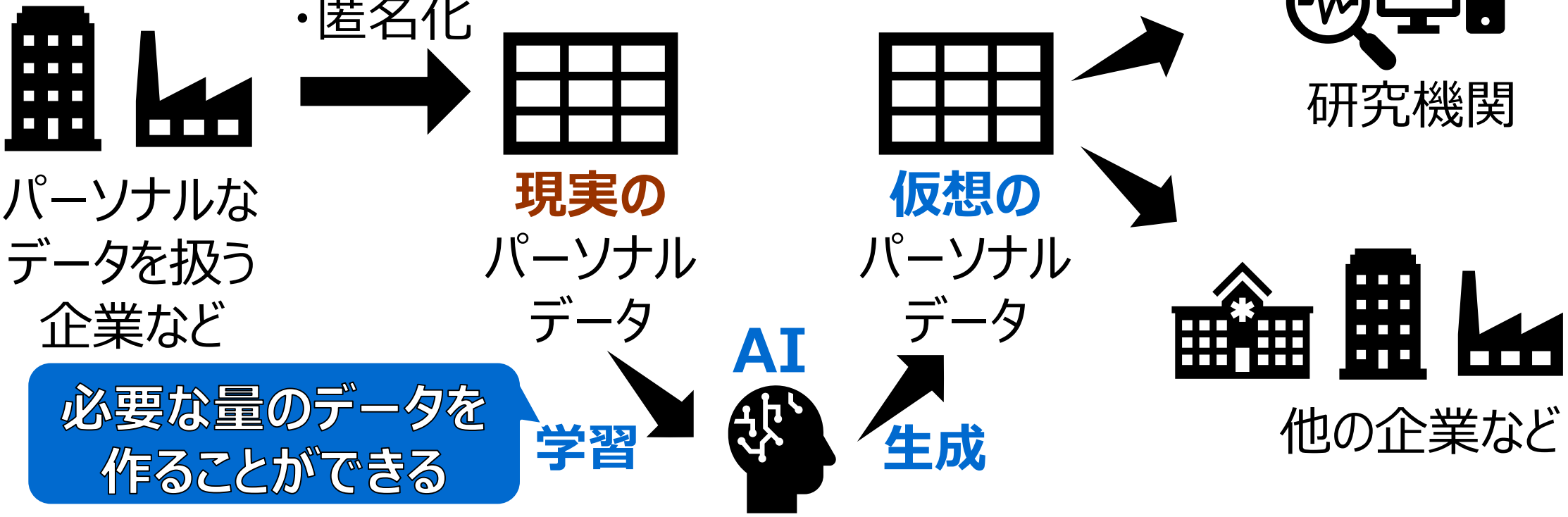
*A. Narayanan, et al., IEEE Symp. on Security and Privacy (2008)

新技術説明会
New Technology Presentation Meetings!

データ連携基盤整備の課題

❶プライバシー保護の必要性
❷十分なデータ量が得られない

非実在データのため
共有のリスクが低い



必要な量のデータを
作ることができる

現実のデータに極めて類似した特徴をもった
仮想のデータを生成することで解決できる

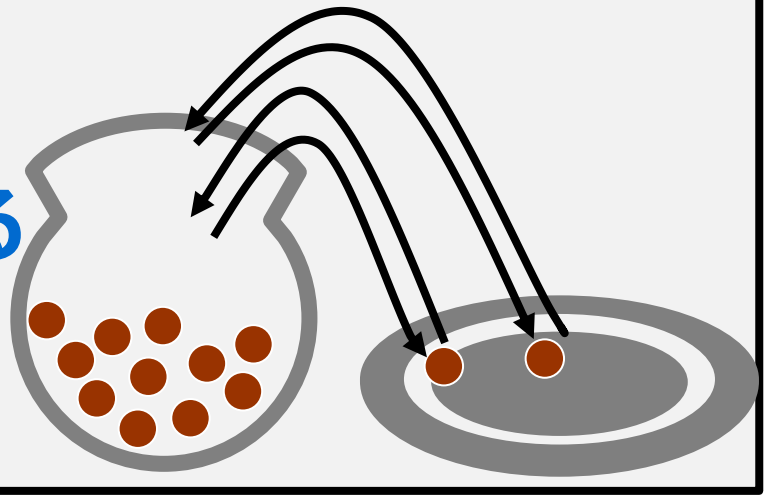
従来技術❶　AI非活用

ブートストラップリサンプリング

- 観測データの再活用（復元抽出）

👍 現実のデータの性質を維持できる

👎 未観測データを生成できない

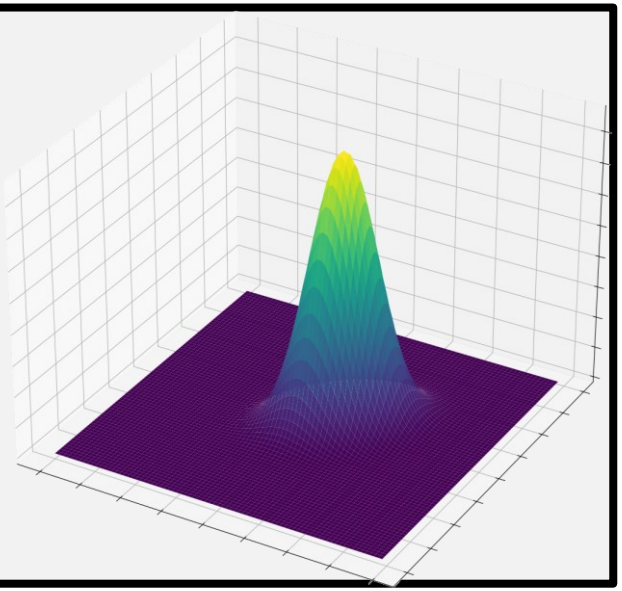


多変量正規分布

- 分布を仮定した生成

👍 未観測データを生成できる

👎 現実のデータの性質再現に限界あり



AIを活用した高度なデータ生成手法への期待が高っている

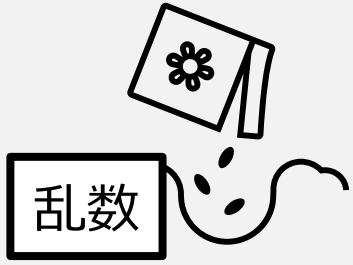
新技術説明会
New Technology Presentation Meetings!

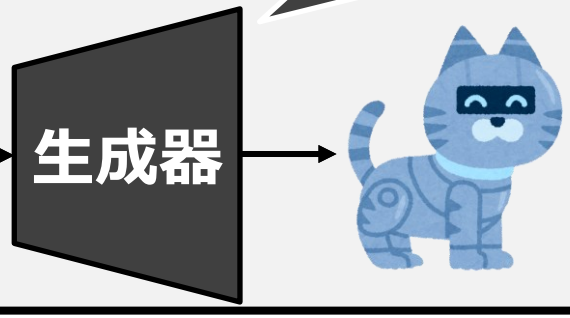
従来技術❷　AI活用

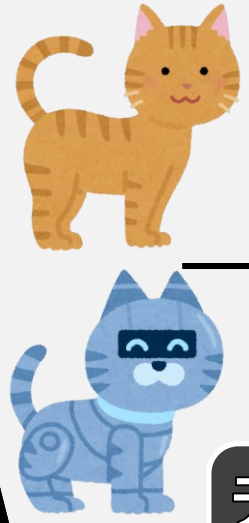
敵対的生成ネットワーク(GAN)

👍生成速度が速い
👎学習が安定しない

乱数からデータを生成するAI





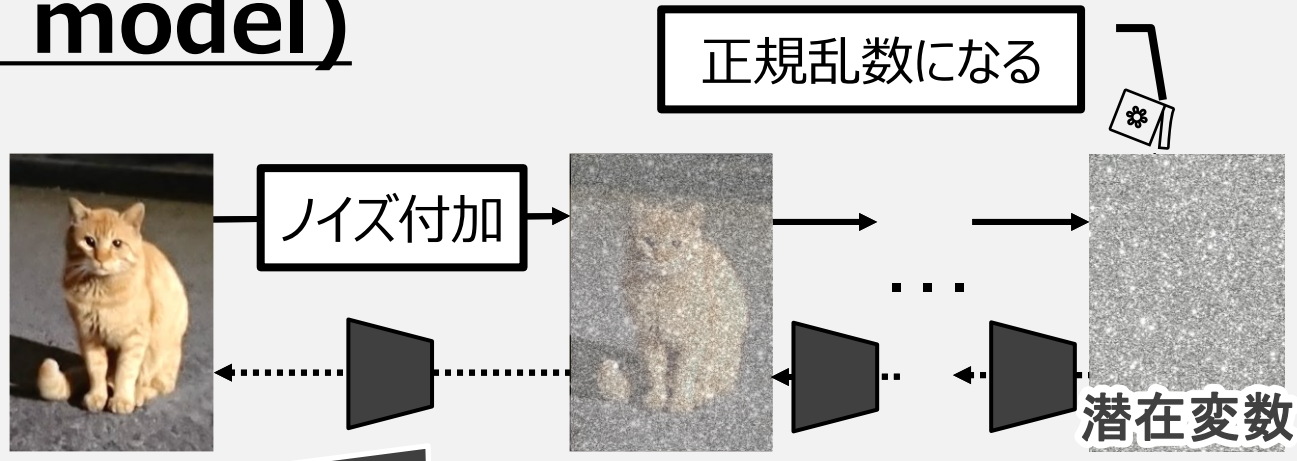


REAL
or
FAKE

データが本物か偽物かを予測するAI

拡散モデル(Diffusion model)

👍学習が安定している
👎生成速度が遅い



ノイズ除去を通してデータを生成するAI

従来技術の問題点

❶AI非活用手法の課題

- 現実のデータの性質を維持しつつ
  未観測のデータを生成することが難しい

❷AI活用手法の課題

- 学習したデータに従う分布や特徴を持った
  仮想データしか生成できない
- 人間が明示できない特徴に基づく任意のデータ群を
  再現した仮想データの生成には再学習が必要である

新技術の特徴

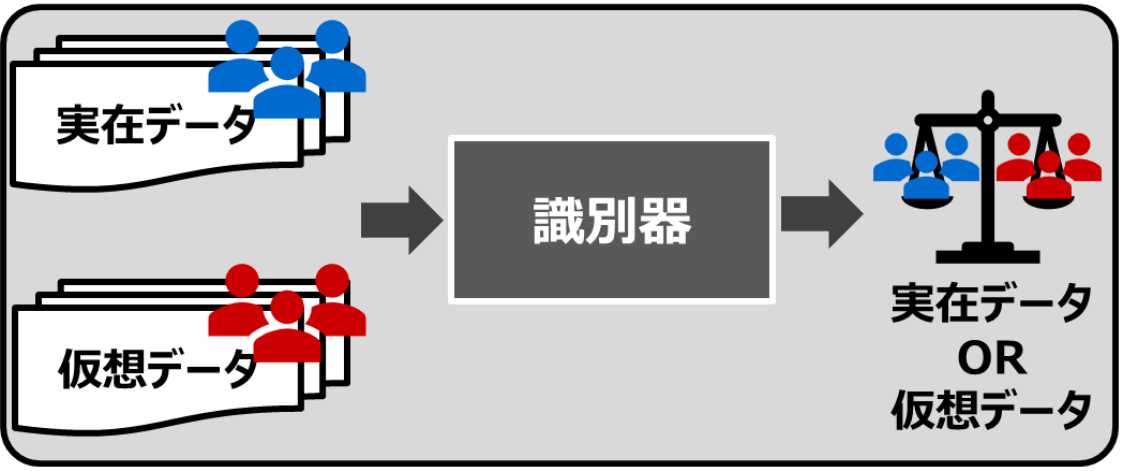
- AIを活用することで、**現実のデータの性質を維持しながらも実在しない（未観測の）仮想データ**の生成を実現した

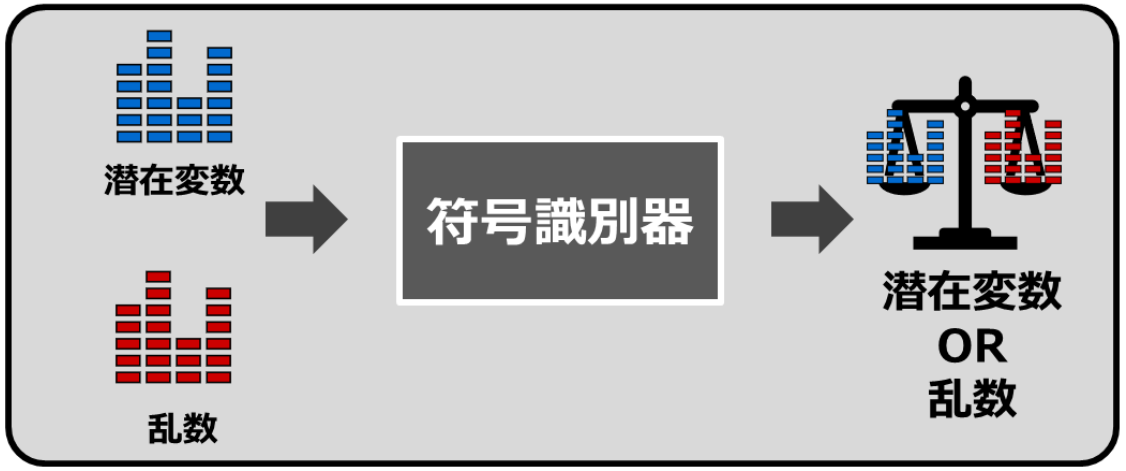
- 符号化器を導入することで種となる潜在変数（乱数）と生成するデータの関係を明示的に学習可能にした

- 符号化器を活用することで**任意の特徴を持つデータ群を再学習不要**で狙って生成可能にした

新技術の特徴

- ✓潜在変数と生成するデータの間の関係を明示的に学習する符号化器
- ✓任意の特徴を持つデータ群から潜在変数を推定することで、再学習なしに特徴を模倣した仮想データ生成





結果の例

食習慣と身体状況に基づく肥満度推定のためのデータセット

- メキシコ、ペルー、コロンビアの人々の食習慣と身体状態のデータで構成される
- ✓ 1,688人分のデータをAIの学習に使用し、学習後のAIで仮想データを生成した

離散値	種類数
性別	2
肥満の家族歴	2
高カロリーな食事	2
間食の頻度	4
喫煙の有無	2
摂取カロリーの管理	2
飲酒頻度	4
交通手段	5
肥満度	7

連続値
年齢
身長
体重
野菜摂取頻度
食事回数
水を飲む量
1週間の運動頻度
1日の情報機器使用時間

評価方法

・データの概観

現実のデータと仮想データをUMAPで次元削減した空間で比較した

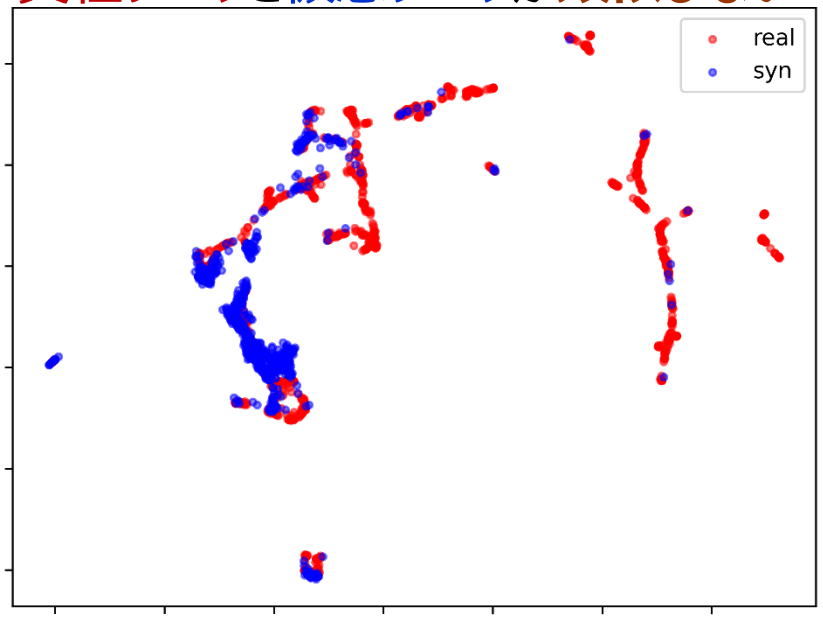
・属性間の相関関係

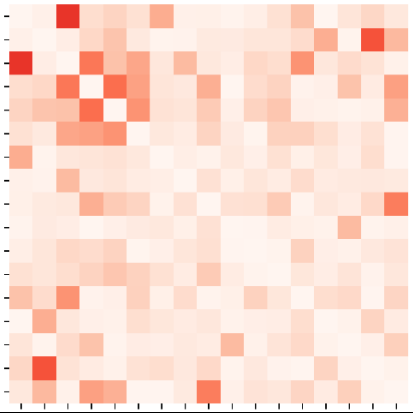
現実のデータの相関と仮想データの相関の差分を一対比較した

結果の例

CTGAN
[L. Xu+, NeurIPS, 2019]

UMAPにより次元削減した空間で
実在データと仮想データが**類似しない**

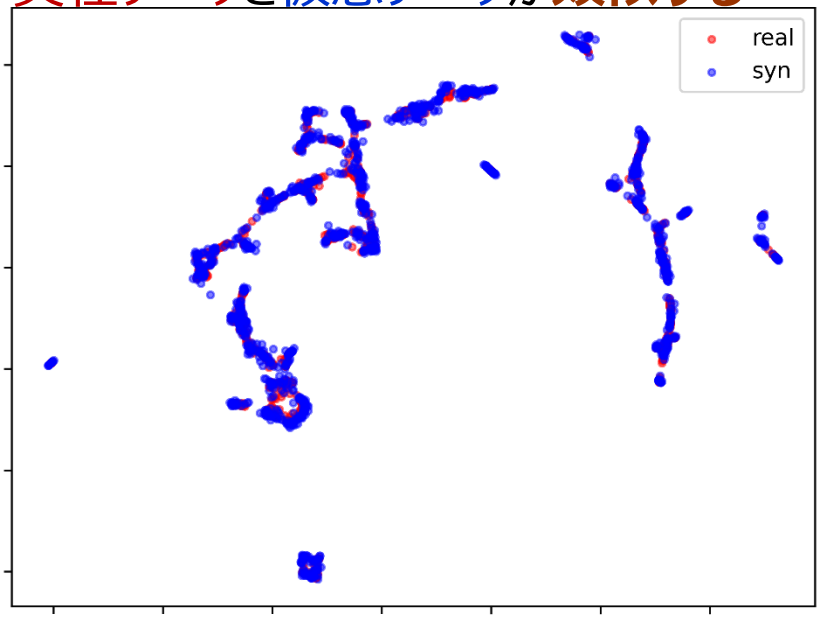


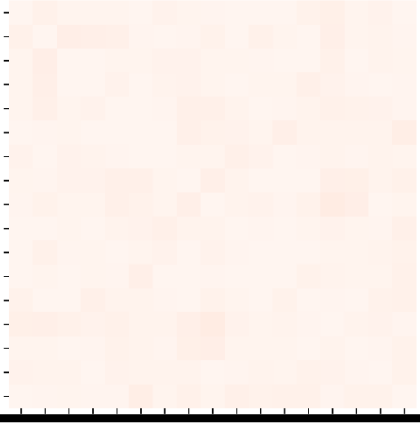


実在データの相関
と
仮想データの相関
の**差分が大きい**

開発技術

UMAPにより次元削減した空間で
実在データと仮想データが**類似する**



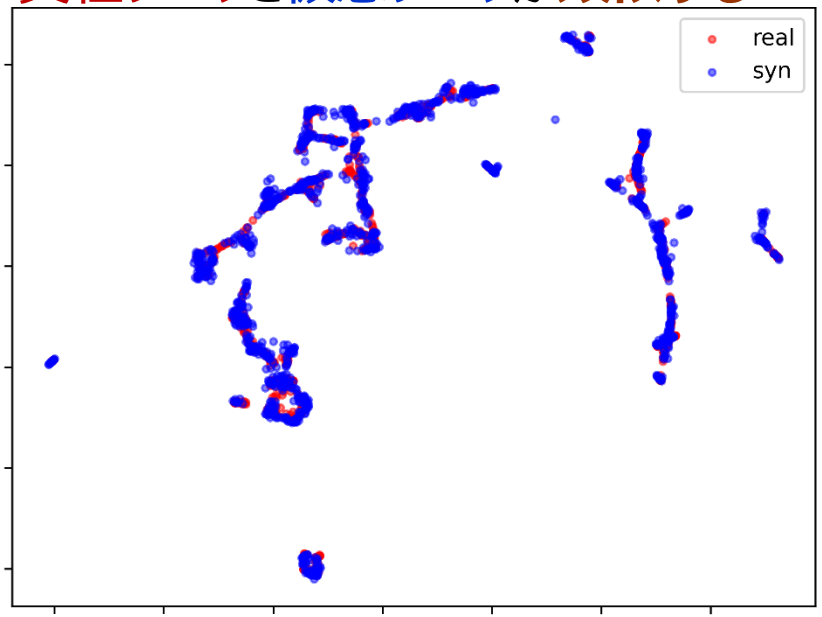


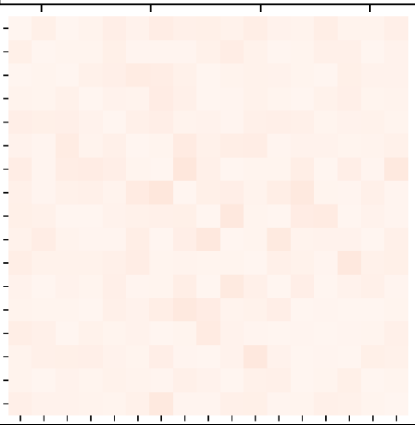
実在データの相関
と
仮想データの相関
の**差分が小さい**

結果の例

TabDDPM
[A. Kotelnikov+, PMLR, 2023]

UMAPにより次元削減した空間で
実在データと仮想データが**類似する**

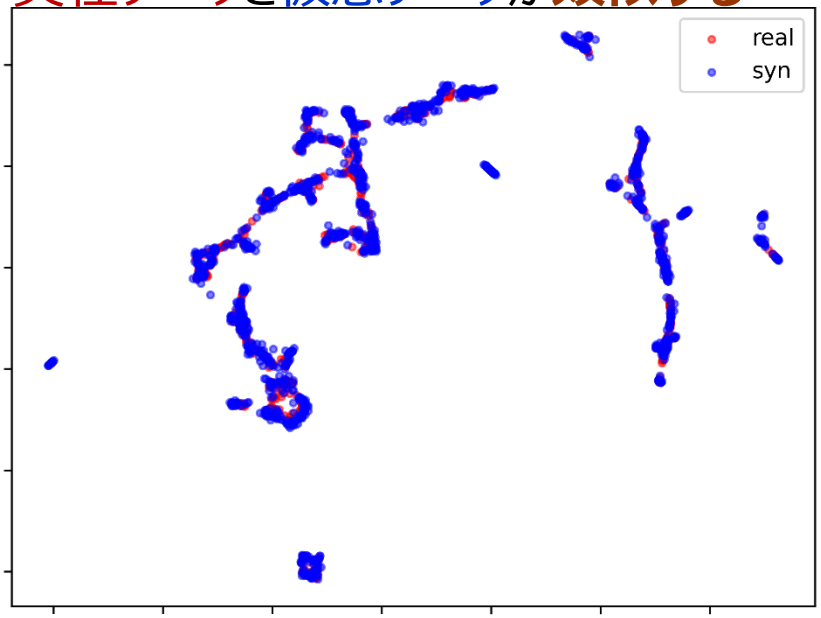


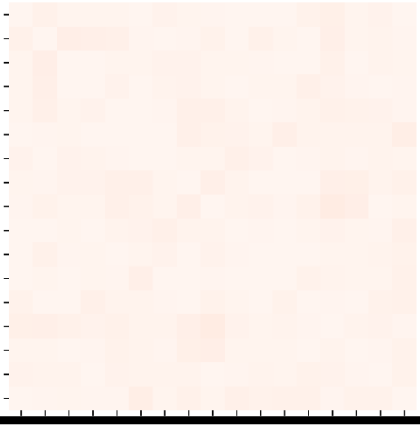


実在データの相関
と
仮想データの相関
の**差分が小さい**

開発技術

UMAPにより次元削減した空間で
実在データと仮想データが**類似する**





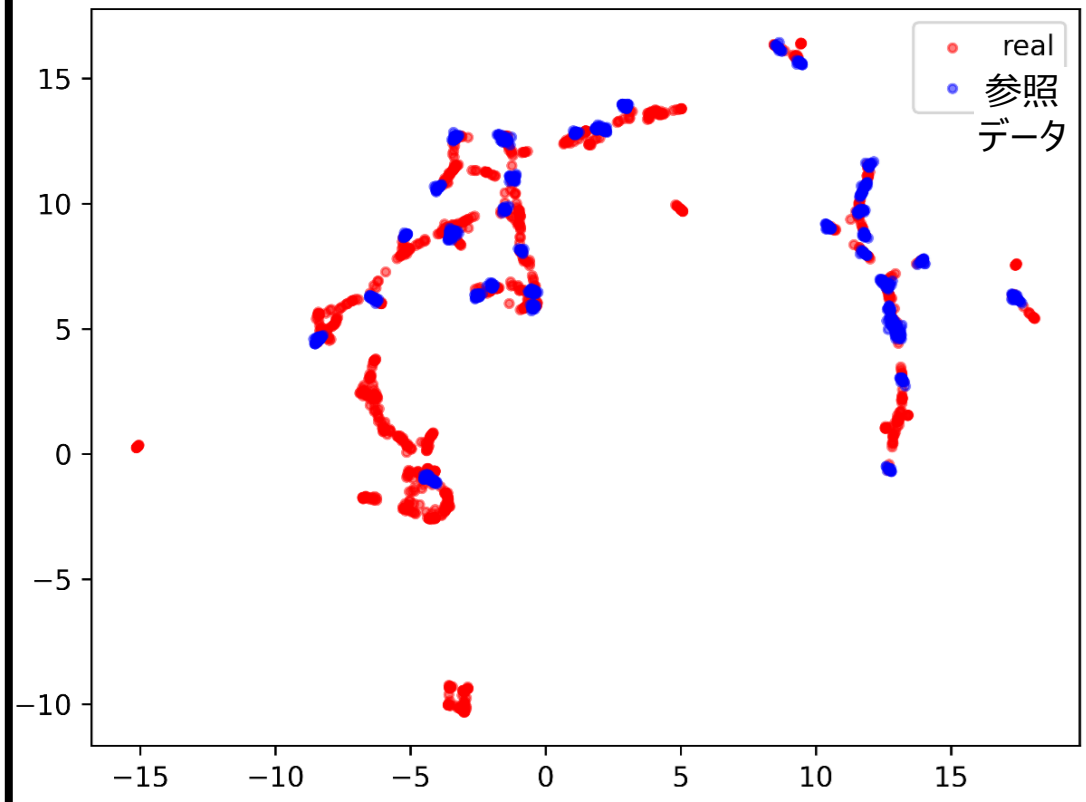
実在データの相関
と
仮想データの相関
の**差分が小さい**

新技術説明会
New Technology Presentation Meetings!

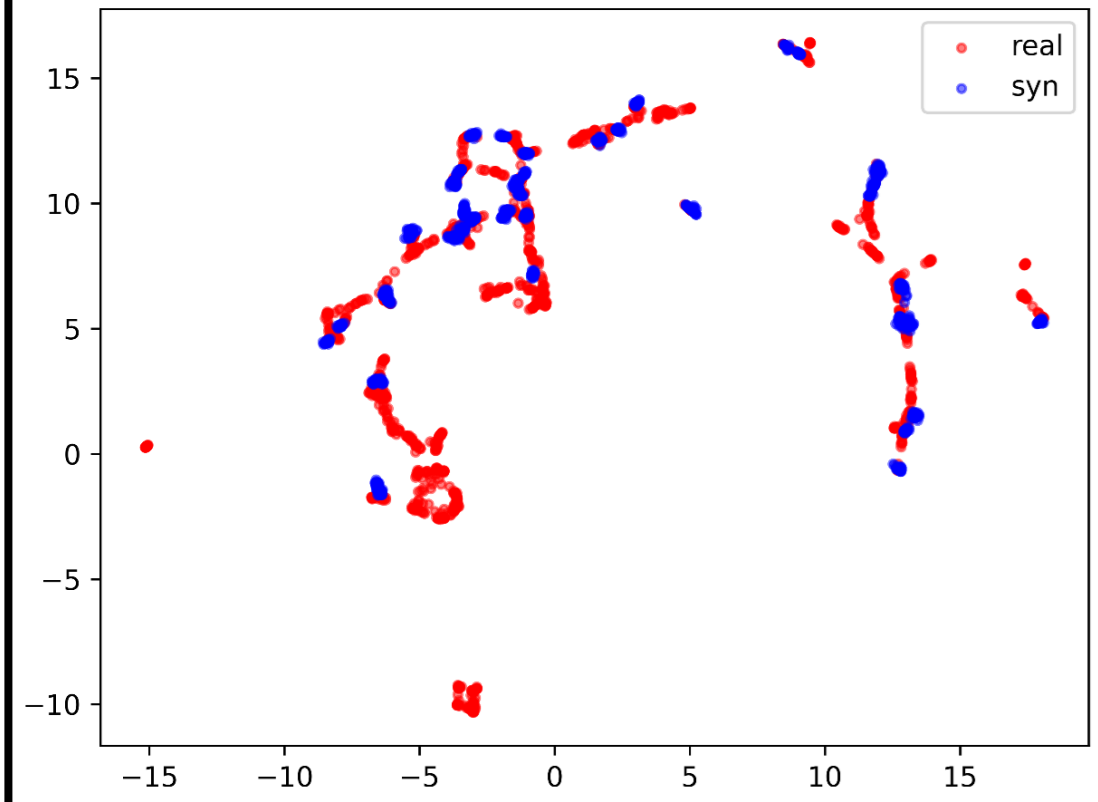
開発技術が可能にした方法の検証

結果の例

現実のデータからランダムに50件のみ抽出した参照データ



開発技術



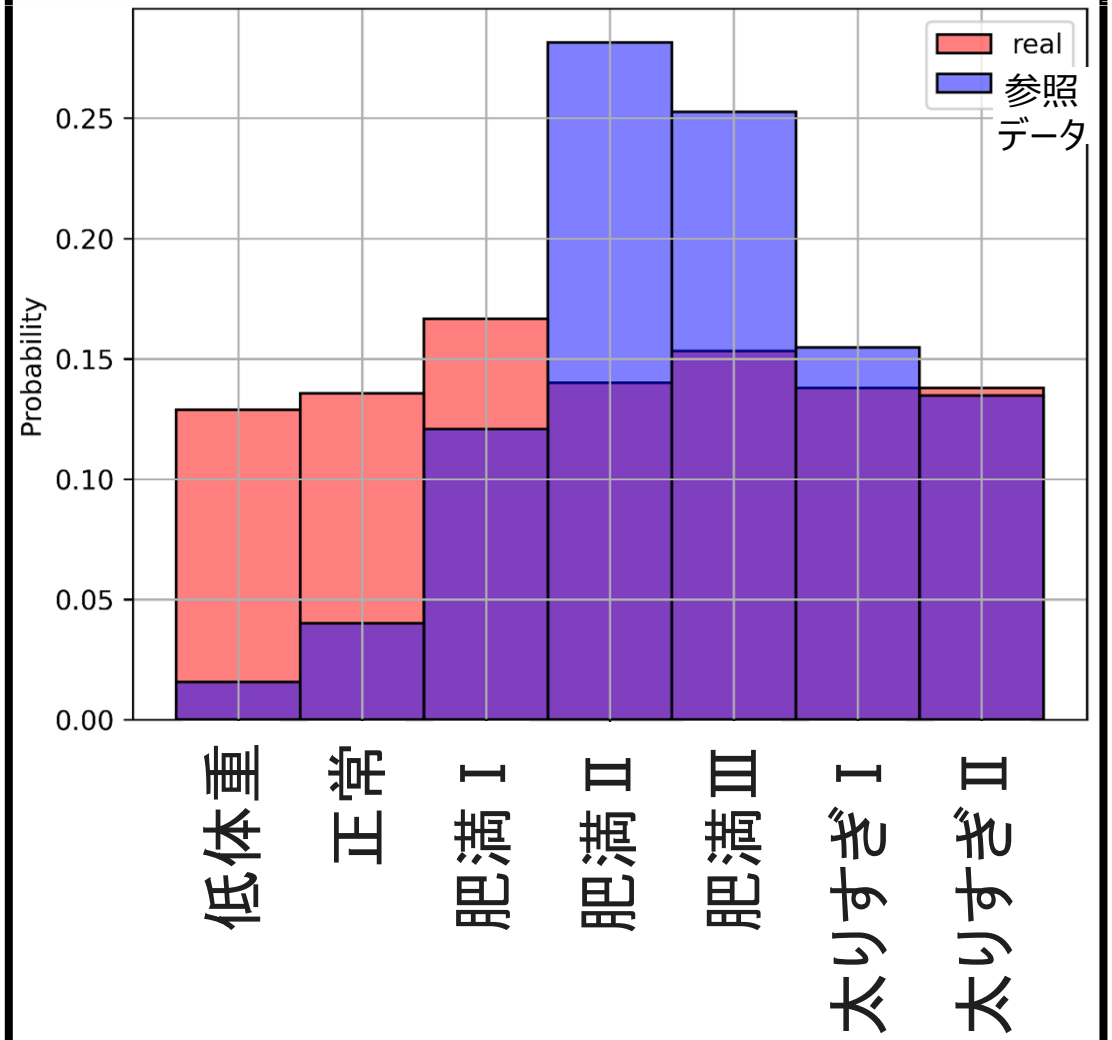
本物のデータと開発技術で再現した仮想のデータは
次元削減した空間上で概ね同様の位置に分布した

新技術説明会
New Technology Presentation Meetings!

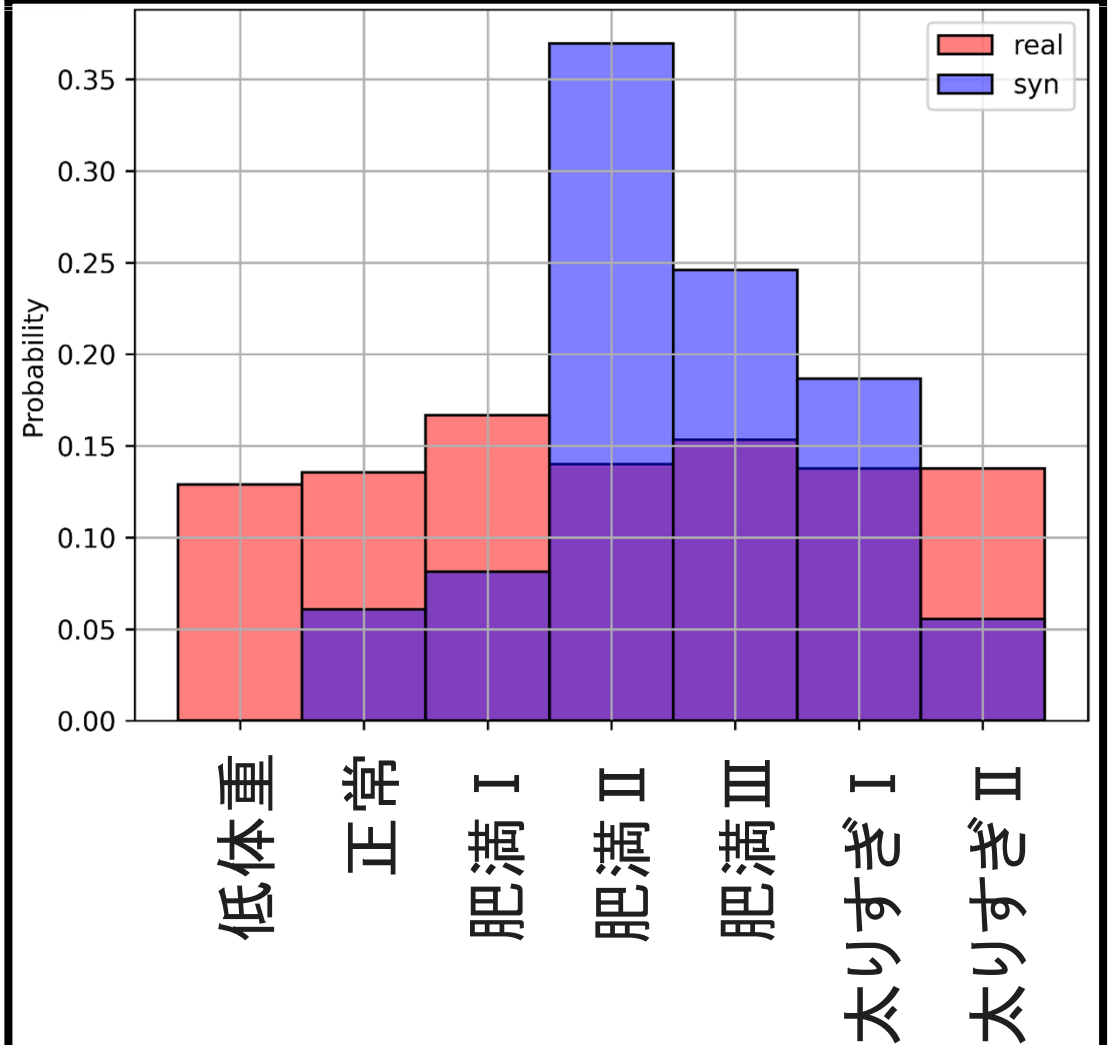
開発技術が可能にした方法の検証

結果の例

現実のデータからランダムに50件のみ抽出した参照データ



開発技術



少量の参照データの特徴を捉えて再現できている

想定される用途

データ連携基盤整備の課題
❶プライバシー保護の必要性
❷十分なデータ量が得られない

- 匿名化処理の代替

- 製薬業界、マーケティング業界など
  パーソナルデータを扱う領域での
  データ共有や解析

- 機械学習、深層学習における学習データ増強

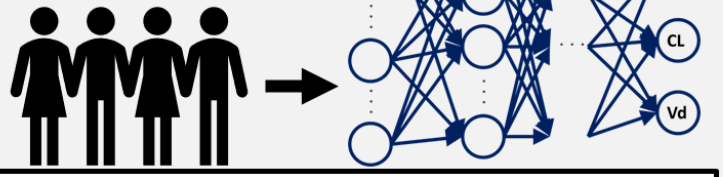
新技術説明会
New Technology Presentation Meetings!

仮想患者データの応用の例
（新技術その２）

機械学習の利点である特徴抽出方法の自動獲得を取り入れつつ、「ブラックボックス化抑制」「経時的な予測」を実現した薬物血中濃度予測モデル

AIモデル

患者データ



ニューラルネットワーク(ANN)

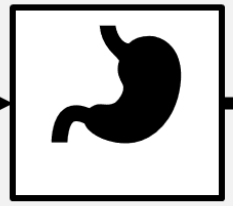
薬物動態パラメータ

吸収速度定数K_a
分布容積V_d
クリアランスCL

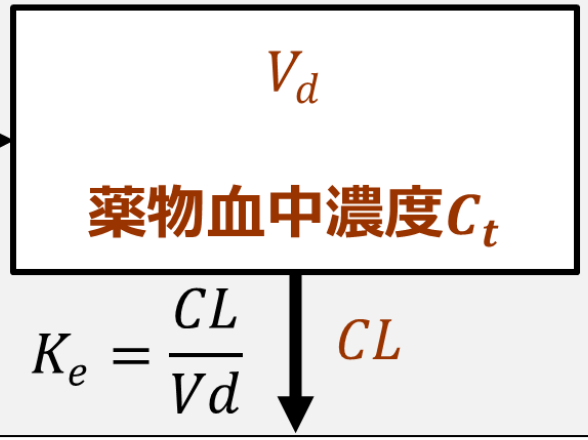
薬学的モデル

投与データ

経口摂取
Dose
Time



K_a



1次吸収過程を含む
1-コンパートメントモデル

薬物動態(PK)モデルとAI(ANN)を組み合わせた**ANN-PKモデル**

・PKモデルとは科学的に導かれた薬物血中濃度を予測する数式である

・AIによる予測の対象をPKモデルでの計算に必要なパラメータにとどめた
➡ブラックボックス化抑制

従来のモデルよりも高精度な薬物血中濃度予測を実現している

新技術説明会
New Technology Presentation Meetings!

仮想患者データの応用の例（新技術その２）

- 薬物血中濃度を含む臨床データは一般に少量である
  - 仮想データをAIの学習に応用することを検討した

❒使用データ

薬剤名	シクロスポリン（免疫抑制剤）
学習データ	80ポイント（34名）
検証データ	9ポイント（9名）

❒CLに対する既知の影響因子

***CL*と年齢**	負の相関
***CL*と体重**	正の相関

AIにこの傾向が確認できれば
科学的妥当性があると判断できる

・現実の臨床データのみ

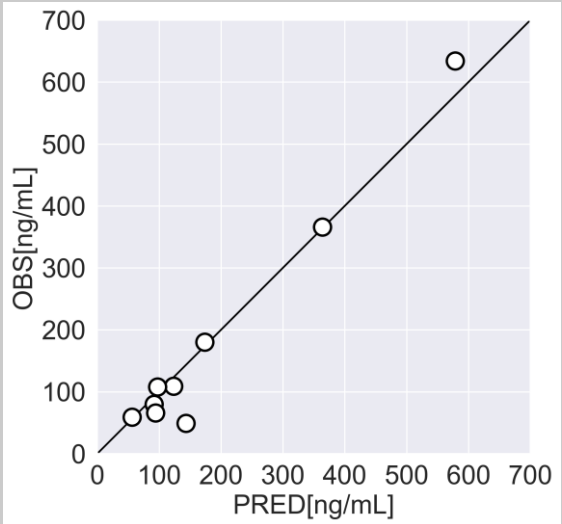
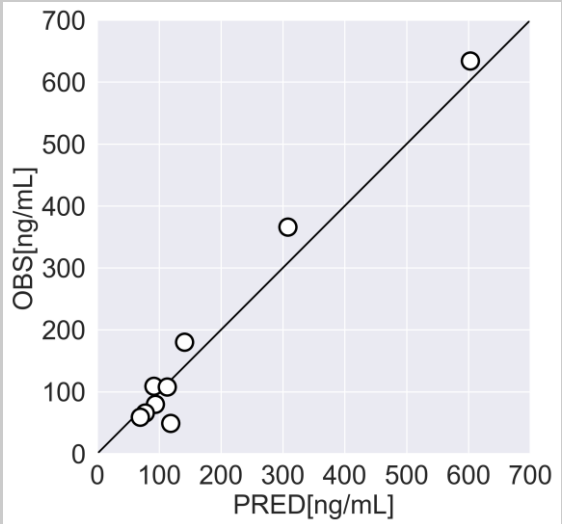
・現実の臨床データと
　合成臨床データの混在

の2条件で
ANN-PKモデルを学習

薬物血中濃度予測精度や予測傾向の妥当性を比較評価

新技術説明会
New Technology Presentation Meetings!

仮想患者データの応用の例
（新技術その２）

	現実のデータのみ	現実のデータ＋仮想データ
		
RMSE [ng/mL]	38.4	35.7
CLと体重の相関	0.95	0.84
CLと年齢の相関	-0.94	-0.95

700
600
500
400
300
200
100
0
OBS[ng/mL]
0 100 200 300 400 500 600 700
PRED[ng/mL]

仮想データをAIの学習に
応用することで
科学的妥当性を維持しつつ
予測精度の向上を実現

実用化に向けた課題

- 多種多様なデータに対応できるかの検証が不足している
- 生成した仮想データの品質評価が不十分である
- 機械学習や解析の際に、現実のデータと同水準には至っていない

企業への期待

- 現実世界の多様なデータの提供
- 仮想データの多角的な品質評価への協力
- 本技術に対するニーズの紹介や提案

企業への貢献、PRポイント

- パーソナルなデータの扱いに困っていたり、それが障害となっていて事業が停滞している企業に貢献できると考えている
- 様々な状況でデータ不足に悩んでいる企業に貢献できると考えている
- 導入にあたってのAI技術指導等ができる

新技術説明会
New Technology Presentation Meetings!

本技術に関する知的財産権

- **発明の名称**：データ生成装置及びデータ生成方法
- **特許番号**　：特願2023-039932、特開2024-130290
- **出願人**　：学校法人日本大学
- **発明者**　：関弘翔、辻泰弘、細野裕行、宮野咲紀、若月（旧姓 尾上）知佳

- **発明の名称**：薬物血中濃度予測装置、薬物血中濃度予測プログラム及び薬物血中濃度予測方法
- **特許番号**　：特許第7462182号
- **出願人**　：学校法人日本大学、国立大学法人九州工業大学
- **発明者**　：辻泰弘、河野英昭、尾上知佳、細野裕行、関弘翔

産学連携の経歴

- 2015年-2017年　電気電子系企業A社からの受託研究実施
- 2015年-2022年　JSTとJICA連携のSATREPSプロジェクトに参加
- 2021年-2022年　(公財)カシオ科学振興財団の研究協賛に採択
- 2023年-2024年　(公財)天野工業技術研究所の研究助成に採択
- 2024年-2026年　科研費若手研究に採択
- 2024年-　製薬企業B社にAI技術指導実施

お問い合わせ先

Nubic 日本大学産官学連携知財センター
ＴＥＬ　03-5275-8139
ＦＡＸ　03-5275-8328
E-mail　nubic@nihon-u.ac.jp