2020/05/09 JAMI 関西・KMI講演会

多様な医療データ活用のための機械学習技術

小島　諒介
（京都大学　医学研究科）

講演資料：http://small-island.work/

もくじ

1. 背景：医療リアルワールドデータ

2. テーブルデータに対する機械学習

3. 医療時系列データ

[研究紹介] 血圧変動予測

4. グラフデータに関する研究紹介

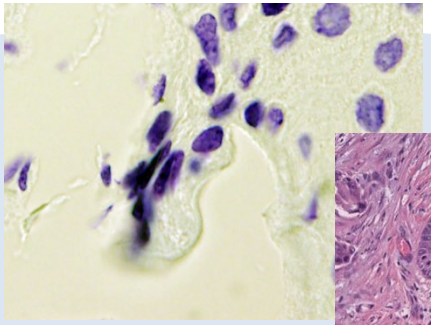
[研究紹介]ゲノム関連知識グラフの構築とその活用

[研究紹介]kGCN： ライフサイエンス向け

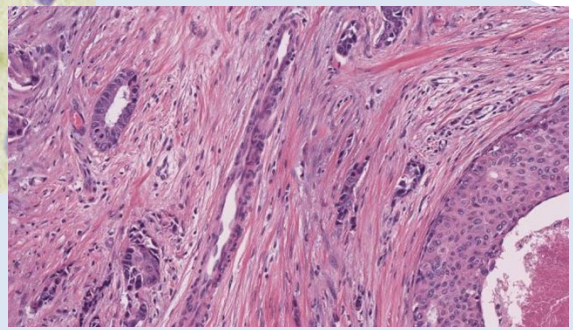
グラフニューラルネットワークツール

多様な医療データ（1/2）

画像

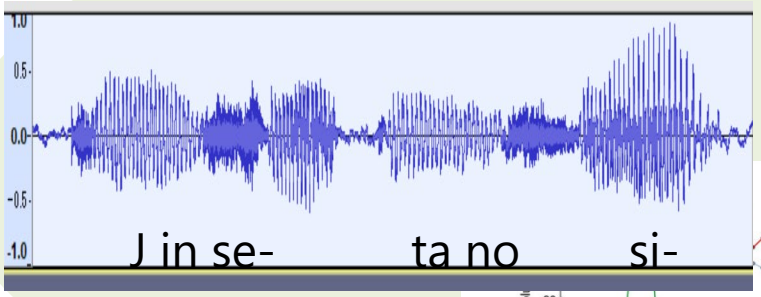


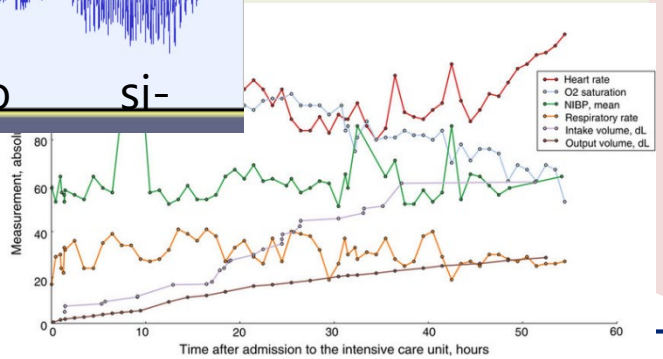
細胞診画像



組織病理画像 (TCGA)

音声・信号データ





心電・脳波・ＩＣＵ

自然言語

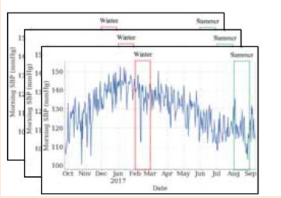
カルテ等
公開文書

MIMIC-III, a freely accessible critical care database

多様な医療データ（2/2）

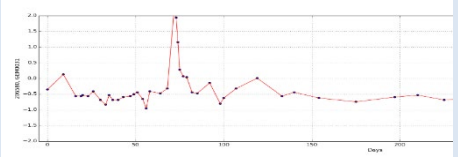
性質の異なるデータが複雑に関係して存在している

収集データ



・実験データ
・研究計画に基づいた収集データ
欠損：少
データ量：少

電子カルテなどの記録データ



・記録データ
欠損：多
データ量：多

公共データ

・薬剤情報
・ゲノム関連情報
データベース間の揺らぎ
アノテーション：少

限りあるデータを有効活用するためには
多様なデータを適切に活用することが求められる

もくじ

1. 背景：医療リアルワールドデータ

2. テーブルデータに対する機械学習

3. 医療時系列データ

[研究紹介] 血圧変動予測

4. グラフデータに関する研究紹介

[研究紹介]ゲノム関連知識グラフの構築とその活用

[研究紹介]kGCN: ライフサイエンス向け

グラフニューラルネットワークツール

テーブルデータに対する機械学習

予測の手掛かりにしたいもの

説明変数

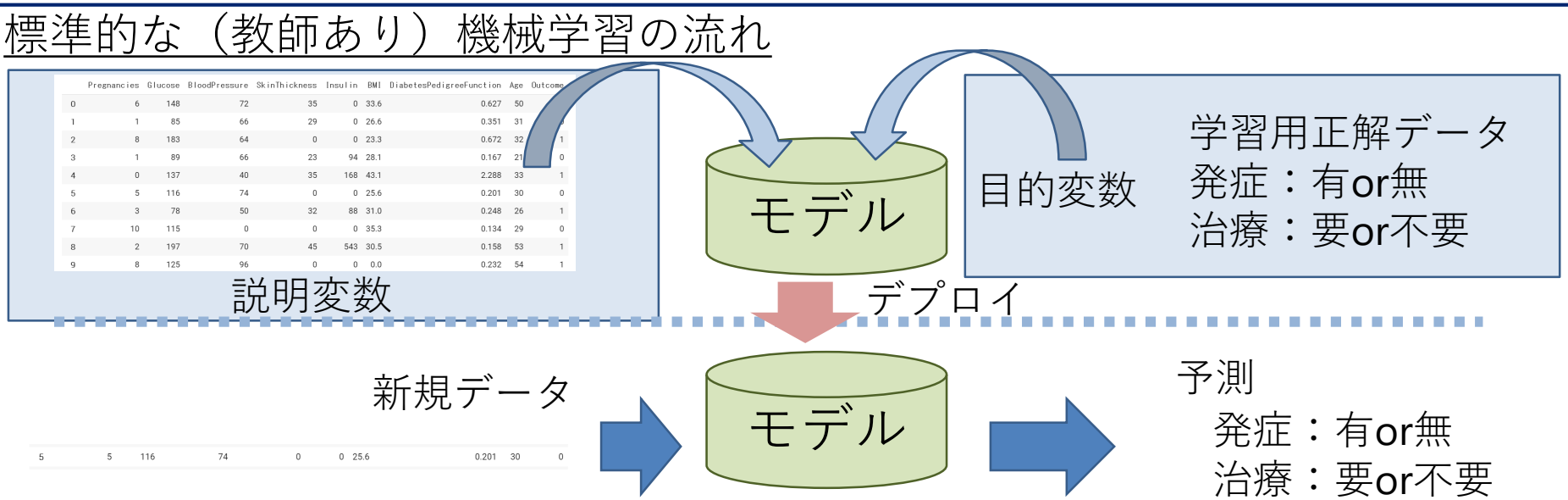
予測したいもの

目的変数

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1
5	5	116	74	0	0	25.6	0.201	30	0
6	3	78	50	32	88	31.0	0.248	26	1
7	10	115	0	0	0	35.3	0.134	29	0
8	2	197	70	45	543	30.5	0.158	53	1
9	8	125	96	0	0	0.0	0.232	54	1

Pima Indiansの糖尿病発症予測データセット

標準的な（教師あり）機械学習の流れ



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1
5	5	116	74	0	0	25.6	0.201	30	0
6	3	78	50	32	88	31.0	0.248	26	1
7	10	115	0	0	0	35.3	0.134	29	0
8	2	197	70	45	543	30.5	0.158	53	1
9	8	125	96	0	0	0.0	0.232	54	1

5	5	116	74	0	0	25.6	0.201	30	0

Automated Machine Learning：上記の流れを自動化してくれる仕組み

- Amazon SageMaker Autopilot
- Google Cloud: AutoML (Tables)
- IBM/AutoAI
- Microsoft/Automated ML
- Sony/Prediction One
- H2O Driverless AI
- Data robot
- PyCaret

画像や自然言語でも同様のサービスは多い

Automated Machine Learning

Pima Indiansの糖尿病発症予測データセットに対する結果 (by PyCaret)

評価尺度

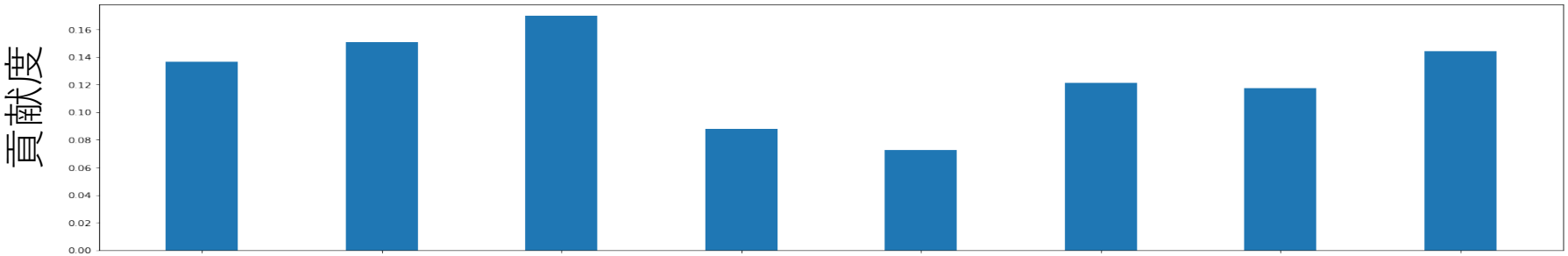
機械学習手法たち

Model	Accuracy	AUC	Recall	Prec.	F1
Ridge Classifier	0.757900	0.000000	0.512900	0.719200	0.594600
Linear Discriminant Analysis	0.756100	0.800900	0.507600	0.716900	0.589900
Logistic Regression	0.743000	0.796000	0.508500	0.680600	0.578000
CatBoost Classifier	0.731900	0.795600	0.518700	0.655300	0.573100
Light Gradient Boosting Machine	0.726400	0.774500	0.557300	0.622800	0.584900
Random Forest Classifier	0.726300	0.745600	0.449100	0.662900	0.532900
Gradient Boosting Classifier	0.717000	0.778000	0.512900	0.620300	0.555500
Extra Trees Classifier	0.717000	0.773000	0.464900	0.633200	0.531000
Ada Boost Classifier	0.715300	0.774300	0.504100	0.610000	0.546200
Extreme Gradient Boosting	0.707700	0.772700	0.508200	0.598800	0.544600
K Neighbors Classifier	0.694800	0.721300	0.514000	0.574500	0.536400
Decision Tree Classifier	0.694800	0.669400	0.584500	0.560200	0.568700
Quadratic Discriminant Analysis	0.651800	0.000000	0.000000	0.000000	0.000000
Naive Bayes	0.647900	0.692600	0.089800	0.448100	0.144900
SVM - Linear Kernel	0.588500	0.000000	0.426300	0.442300	0.372500

応用上重要な機能：予測貢献度の表示

応用上では、単純な予測だけでなく、何が予測に貢献しているかを表示したい

Outcomeに対する**変数**の貢献度



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

個別のサンプル（患者）ごとの貢献度 (by SHAP)

応用上重要な機能：予測貢献度の表示

予測貢献度の可視化手法たち
（XAI: Explainable AI という言葉もある）

- Partial Dependence Plots [Friedman+ 2001]
- M-Plots & ALE [Apley+ 2016]
- Individual Conditional Expectation [Goldstein+ 2017]
- Permutation feature importance [Altmann+ 2010]
- Global surrogate
- Local surrogate: LIME [Ribeiro+ 2016]
- SHAP [Lundberg+ 2017]
- Gradient-based methods
- Attention
- Graph-base: GNNExplainer [Ying+ 2019]

もくじ

1. 背景：医療リアルワールドデータ

2. テーブルデータに対する機械学習

3. 医療時系列データ

[研究紹介] 血圧変動予測

4. グラフデータに関する研究紹介

[研究紹介]ゲノム関連知識グラフの構築とその活用

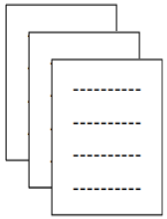
[研究紹介]kGCN： ライフサイエンス向け

グラフニューラルネットワークツール

時系列データ

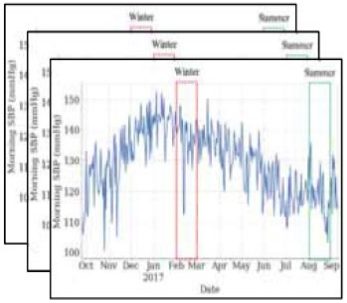
- 時間的な変化を取り扱う必要がある
  - トレンド・季節性・曜日
- 時間的に変化する項目と変化しない項目の２つのモダリティがある

非時系列：
診察室情報



- 既往症
- 体重/身長
- 年齢
- 性別
- ・・・

時系列：
検査値など



- 未来の不確実性
  - １日後の予測、一週間後の予測、一か月後の予測

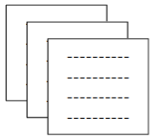
収集データの例：血圧の変動予測

データ：
通院中の患者を対象に423人分の
毎日朝晩の家庭血圧を収集
PREDICT(Prediction of ICT-Home Blood Pressure Variability)
study conducted by Jichi Medical University in Japan

血圧指標と管理基準

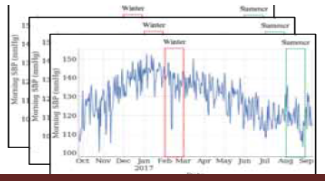
	種別	基準値
平均的な血圧に関する指標	家庭血圧（週間平均で評価）	SBP: 135 / DBP:85 mmHg以上
	診察室血圧 [1]	SBP: 140 / DBP:90 mmHg以上
血圧の変動性に関する指標	診察室訪問間血圧差（Visit-to-Visit [2]	基準は未決定
	日間変動性（Day-by-Day [3]	基準は未決定
	季節変動（Seasonal）	基準は未決定

診察室情報

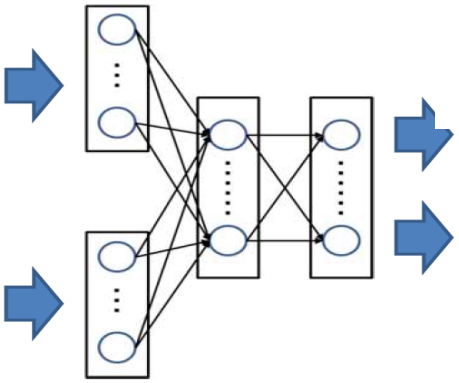


- 既往症
- 体重/身長
- 年齢
- 性別
- ・・・

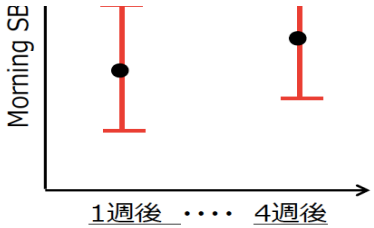
過去56日間の血圧



深層学習モデル



血圧の平均
＋変動性の予測



H. Koshimizu, R. Kojima, K. Kario, Y. Okuno: Prediction of blood pressure variability using deep neural networks. In *International Journal of Medical Informatics*, Vol. 136 pp. 104067, 2020.

入力：時系列データと非時系列データの混合データ
出力：複数時点の同時予測＋変動予測

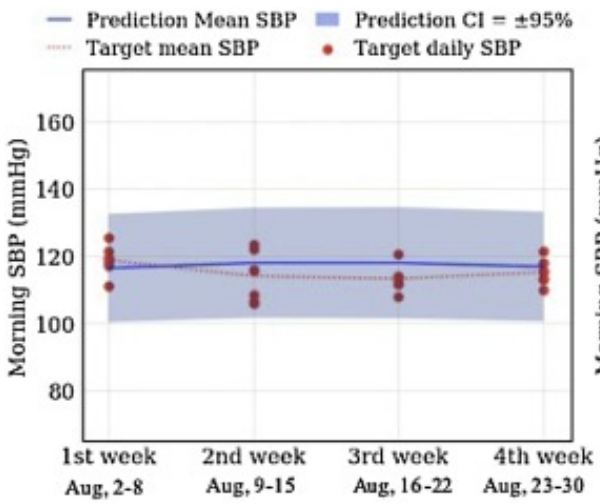
[1] S. Umemura et al., Chap2, The Japanese Society of Hypertension Guidelines for the Management of Hypertension, HYPERTENS RES, 42, pp.1248–1281, 2019.
[2] Rothwell PM et al., Limitations of the usual blood-pressure hypothesis and importance of variability, instability, and episodic hypertension. Lancet.2010;375:938–48
[3] Kuriya M et al., Day-by-day variability of blood pressure and heart rate at home as a novel predictor of prognosis: the Ohasama study. Hypertension, 52(6):1045-50, 2008

結果：血圧の変動予測

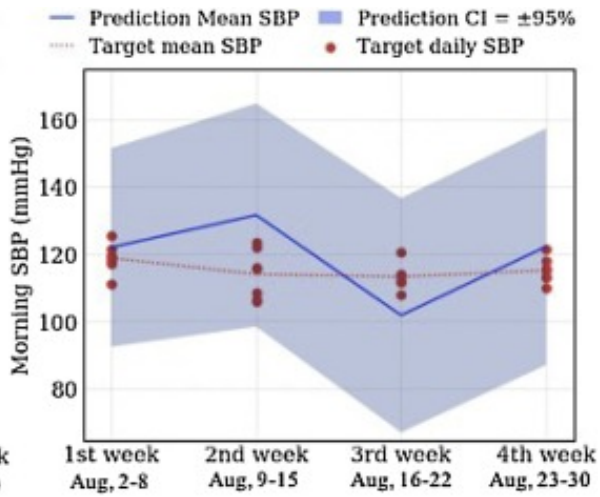
血圧平均値: 5.11-6.65 mmHgで予測可能
血圧変動性: 0.67-0.70

＊電子血圧計の精度基準は
聴診に対して誤差平均5mmHg以内

		RMSE (mmHg)		SR	
予測対象	値	提案手法	ベンチマーク	提案手法	ベンチマーク
1週間後	平均	**5.11**	15.89	**0.69**	0.38
	最良	**5.04**	13.56	**0.70**	0.43
2週間後	平均	**5.91**	16.67	**0.68**	0.37
	最良	**5.82**	13.52	**0.69**	0.43
3週間後	平均	**6.34**	16.43	**0.67**	0.37
	最良	**6.26**	14.07	**0.68**	0.42
4週間後	平均	**6.65**	16.51	**0.67**	0.37
	最良	**6.60**	14.27	**0.67**	0.42



提案法



ベンチマーク

血圧の平均的な値とその変動性（標準偏差）を予測可能性を示した

もくじ

1. 背景：医療リアルワールドデータ
2. テーブルデータに対する機械学習
3. 医療時系列データ

   [研究紹介] 血圧変動予測

4. グラフデータに関する研究紹介

   [研究紹介]ゲノム関連知識グラフの構築とその活用

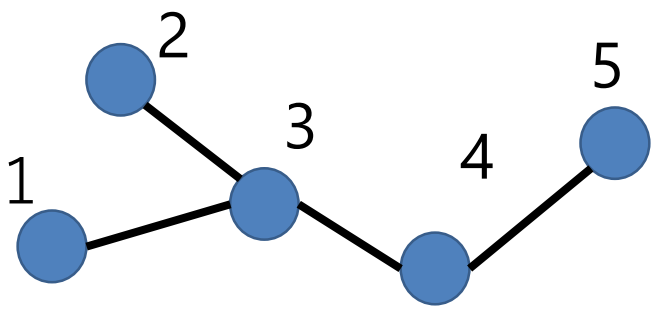
   [研究紹介]**kGCN:** ライフサイエンス向け

   グラフニューラルネットワークツール

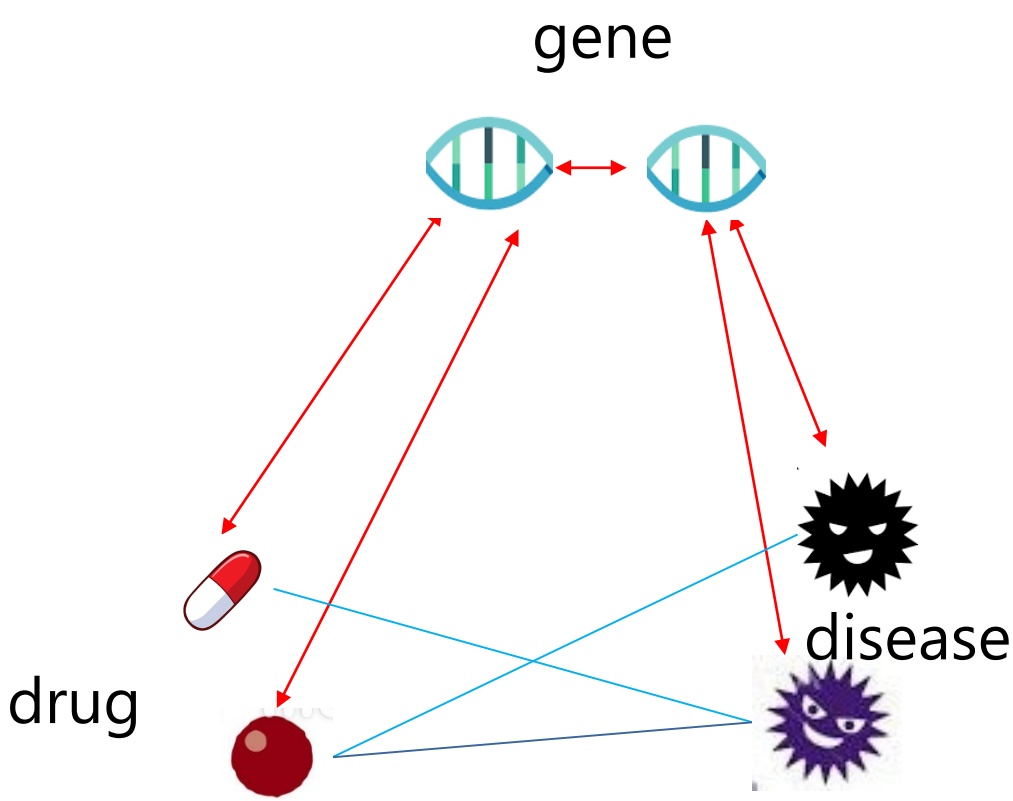
グラフデータ

グラフ:$G = (V, E)$はノードの集合とエッジの集合で構成される

V: ノード（頂点）集合
E: エッジ（辺・リンク）集合



距離は無関係に
つながりだけを
表現している

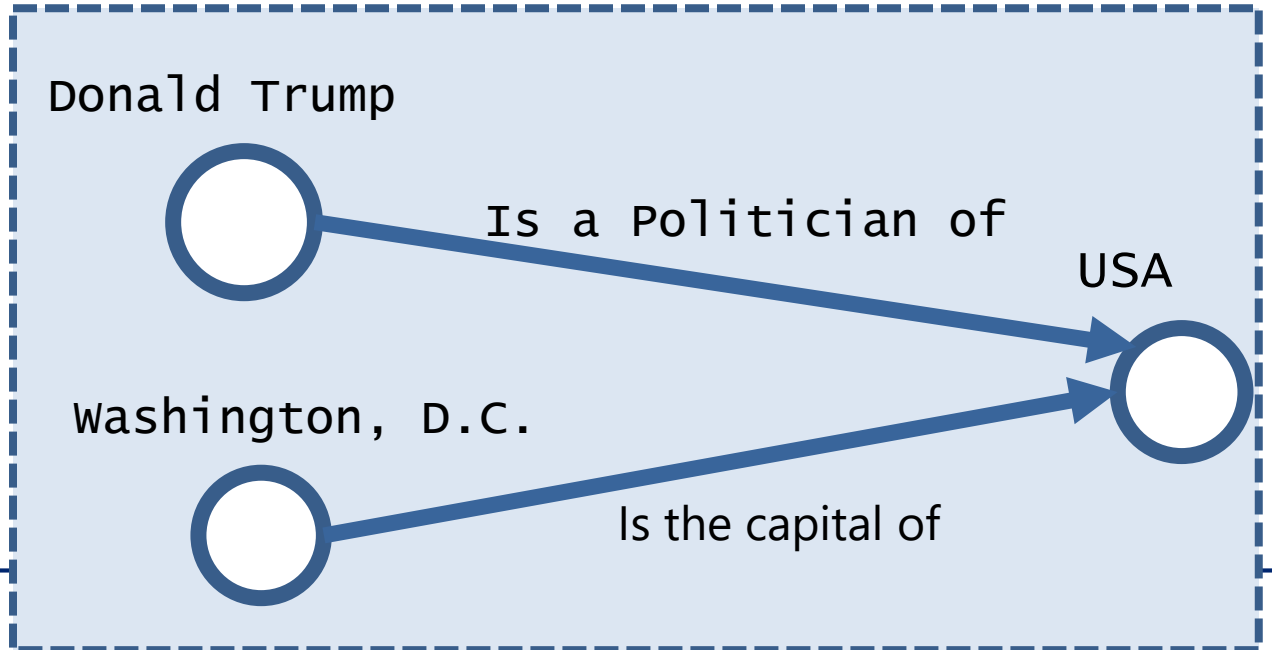


Guney, Emre, et al. "Network-based in silico drug efficacy screening." *Nature communications* 7 (2016): 10331.

知識グラフ

h	ℓ	t
Donald Trump	Is a politician of	USA
Washington, D.C.	Is the capital of	USA

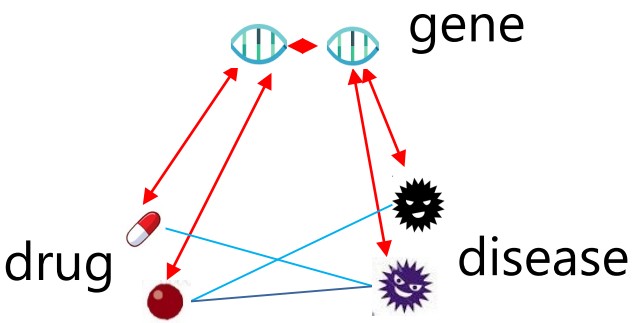
RDF(Resource Description Framework)
Webなどで多数の人が知識を共有する枠組み

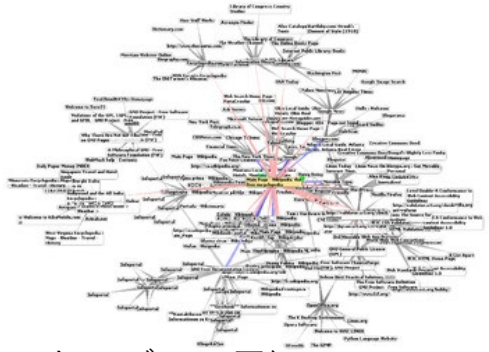


世の中のグラフデータ

- 地図・インフラのネットワーク
- 友達関係などのネットワーク
- 距離から作られるグラフ
（ポイントクラウド・分子中の原子位置）
- 概念・意味の構造
- 化合物の構造式

知識グラフ

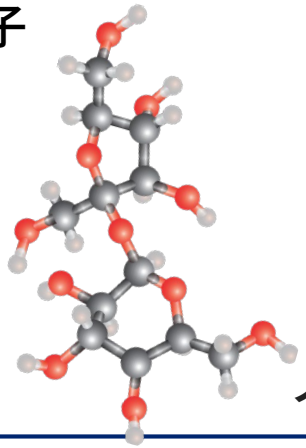


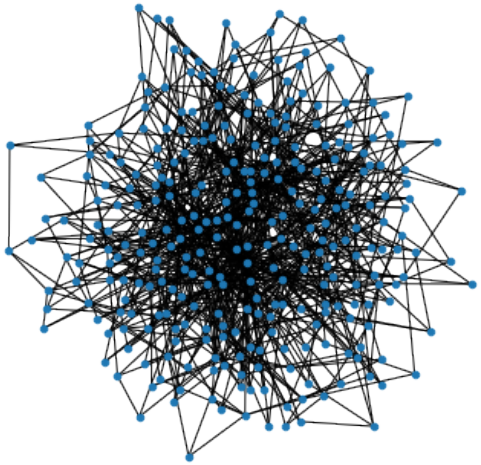


ウィキペディア周辺のWWWの構造

https://ja.wikipedia.org/wiki/%E8%A4%87%E9%9B%91%E3%83%8D%E3%83%83%E3%83%88%E3%83%AF%E3%83%BC%E3%82%AF

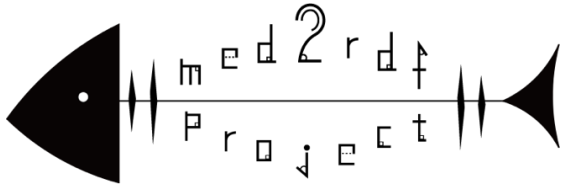
分子





ノード数300のランダムグラフ

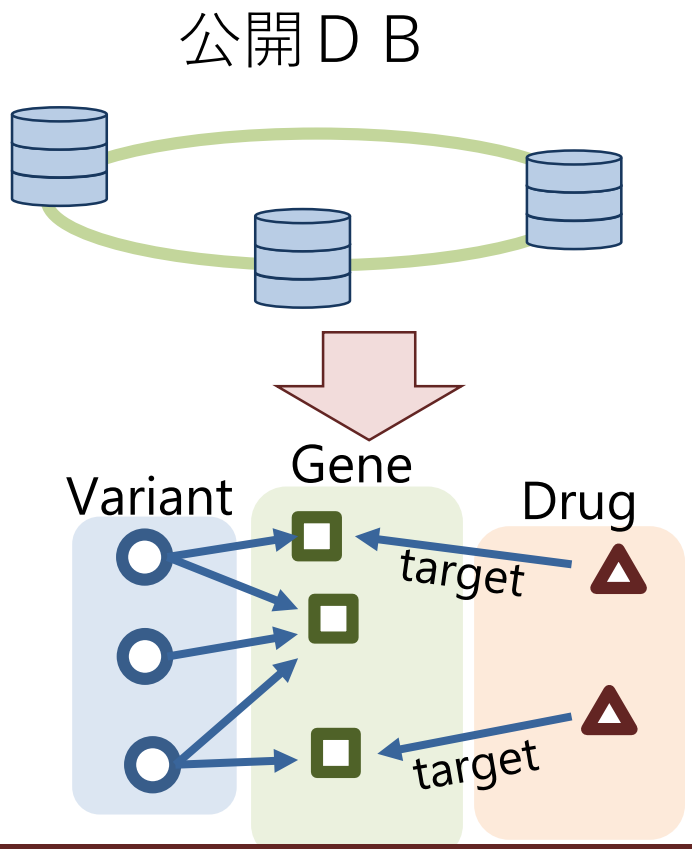
ライフサイエンス知識グラフ：Med2RDF



http://med2rdf.org/

M. Kamada, T. Katayama, S. Kawashima, R. Kojima, M. Nakatsui, Y. Okuno: *Construction of Knowledge-base for Clinical Interpretation of Genomic Variants*.Semantic Web Applications and Tools for the Life Sciences (SWAT4LS), 2017.

ゲノム関連知識グラフを公開DBから構築

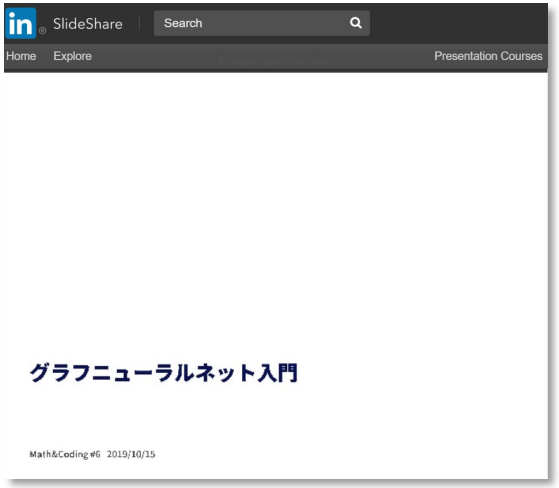


Category	DB	Information
Pathogenicity of variant	ClinVar	Pathogenicity of variant
	COSMIC	Impact of variants in human cancer
	CIViC	Variants with published clinical evidence for cancer
Frequency in population	dbSNP	Genetic variation with MAF
	dbVar	Large scale genomic variation
	ExAC	Allel frequency in epoch population
Effect of variant on protein function	dbNSFP	Effect scores prediction by 21 bioinformatics tools
Drug interaction	DGIdb	Gene-drug interaction
Molecular interaction	HiNT	Protein- protein interaction from interactome resources
	INstruct	3D-protein interaction with structure

複数のデータベース間をグラフとして横断的に見ることができる

グラフニューラルネットワーク

グラフから目的変数を予測できる
ニューラルネットワーク
（説明変数をグラフに置き換えたもの）



PyTorch GeometricでGraph Neural Network（GNN）入門
https://cpp-learning.com/pytorch-geometric/

こちらのブログ記事でも紹介されてました

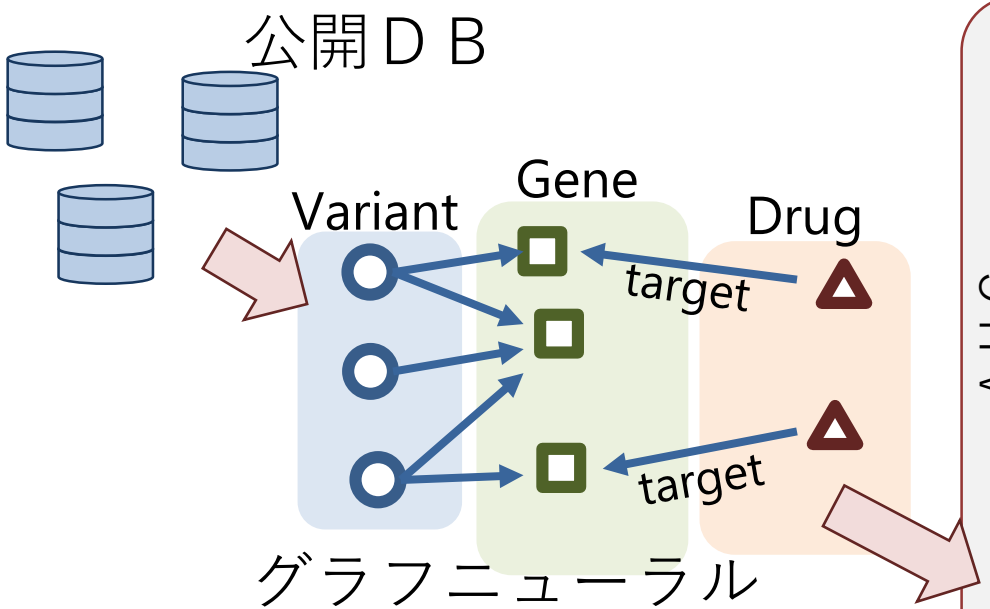
https://www.slideshare.net/ryosuke-kojima/ss-179423718

ライフサイエンス知識グラフ：Med2RDF

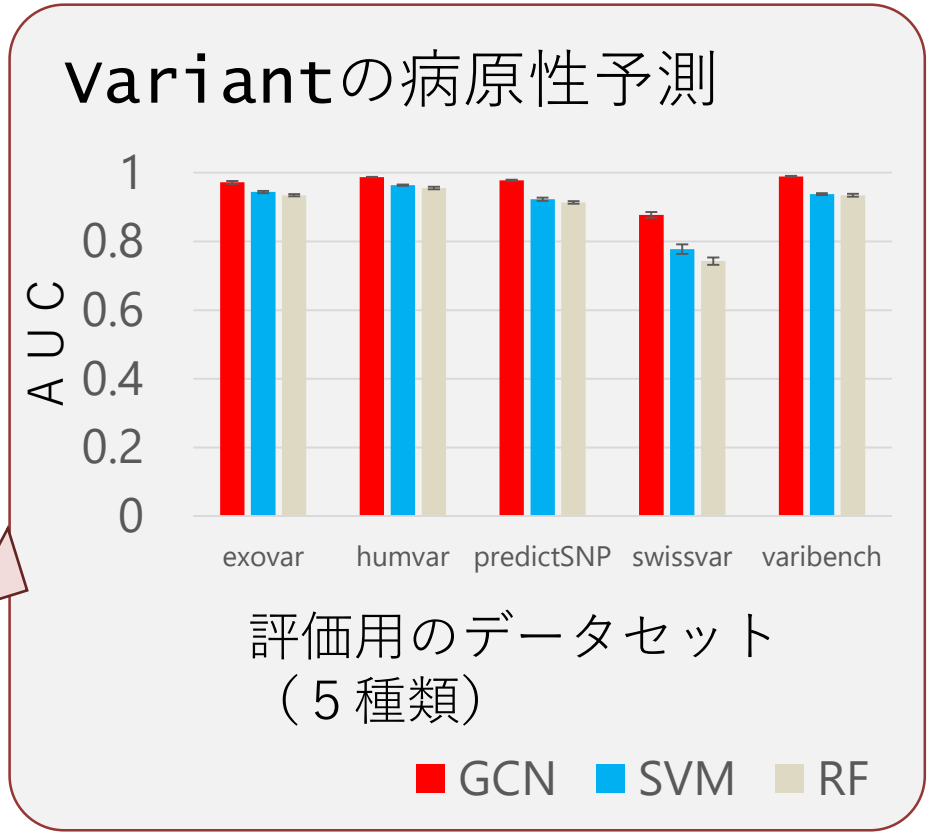
http://med2rdf.org/

M. Kamada, T. Katayama, S. Kawashima, R. Kojima, M. Nakatsui, Y. Okuno: *Construction of Knowledge-base for Clinical Interpretation of Genomic Variants.*Semantic Web Applications and Tools for the Life Sciences (SWAT4LS), 2017.

ゲノム関連知識グラフを公開DBから構築



グラフニューラルネットワークによる予測



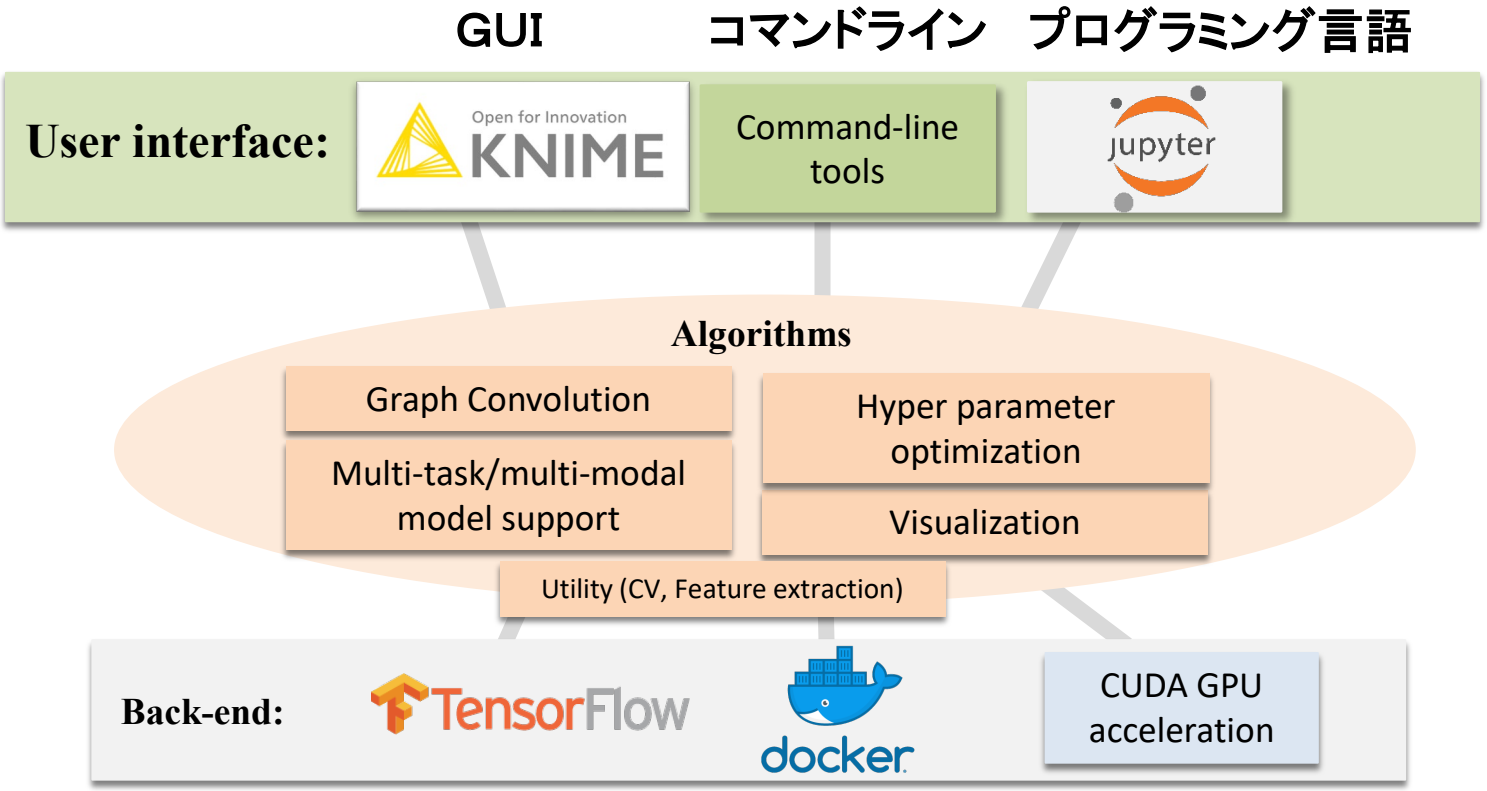
グラフを利用することで、性能を上げることができる

kGCN: Graph Convolutional Network tools for life science

Github: https://github.com/clinfo/kGCN

グラフニューラルネットワーク用ライブラリ

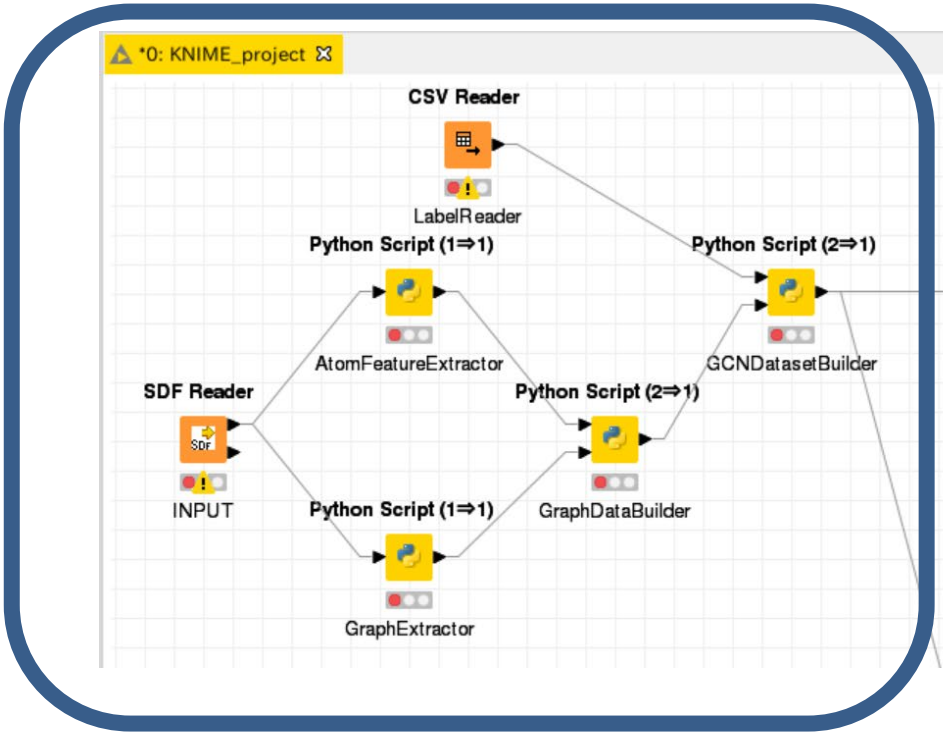
様々な人がグラフニューラルネットワークを含む深層学習を利用可能になることを目指したツール群



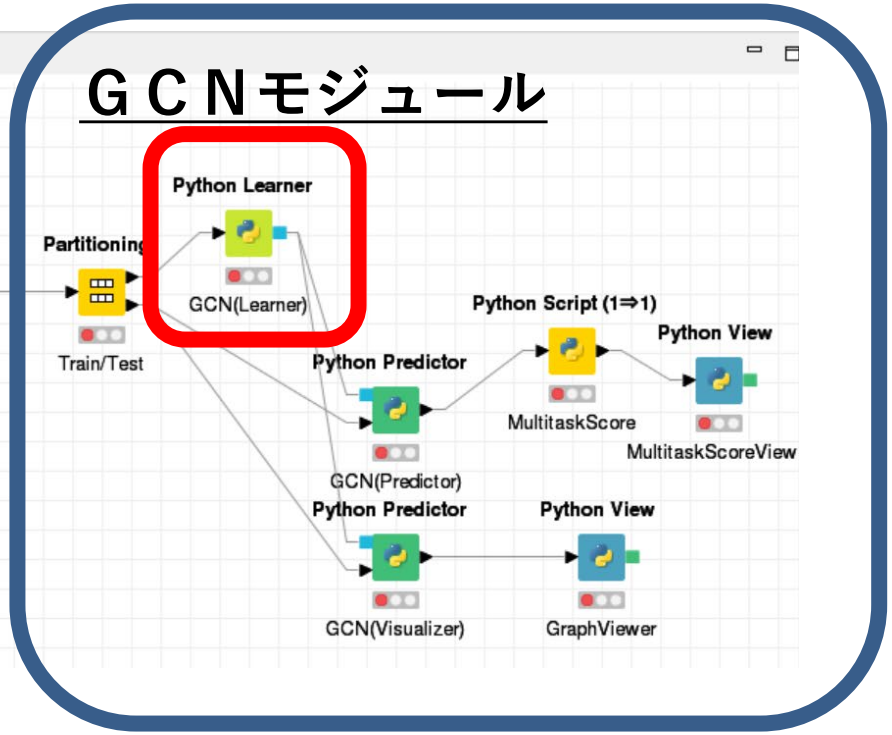
[インターフェース1/3] KNIME インターフェース

ブロックを線でつなげることで、グラフィカルに処理を記述可能

前処理部分



学習・可視化部分



[インターフェース2/3] コマンドライン インターフェース

configファイルに設定を書いて、kgcnコマンドで処理を一括で実行できる

基本的には

学習

```
$ kgcn train --config config.json
```

予測

```
$ kgcn predict --config config.json
```

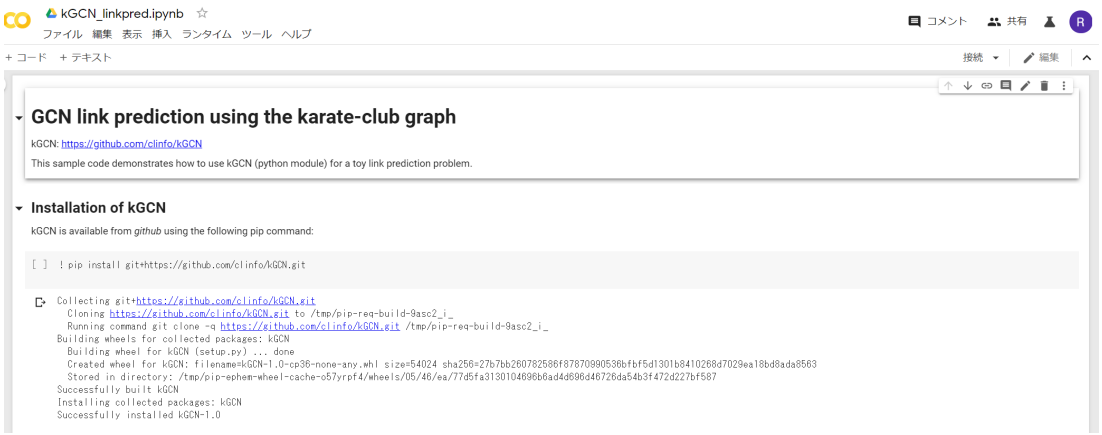
クロスバリデーションを実行

```
$ kgcn train_cv --config config.json
```

[インターフェース3/3] Jupyter notebook

Pip（Pythonの標準的なパッケージツール）でインストール可能、google colaboratoryなどからでも利用可能

https://colab.research.google.com/drive/1E9sCWQKWs2BO_i_zAudne-8nJQigYYLl



基本的には

- グラフを作成
- ニューラルネットワークの構造を定義
- 入力データに関する設定：dictで指定
- 学習に関する設定：dictで指定
- 学習：fit()
- 予測：pred()

応用上重要な機能：予測貢献度の可視化

予測に対してどの入力特徴量が効いているかを可視化する手法

- Partial Dependence Plots [Friedman+ 2001]
- M-Plots & ALE [Apley+ 2016]
- Individual Conditional Expectation [Goldstein+ 2017]
- Permutation feature importance [Altmann+ 2010]
- Global surrogate
- Local surrogate: LIME [Ribeiro+ 2016]
- SHAP [Lundberg+ 2017]
- Gradient-based methods
- Attention
- Graph-base: GNNExplainer [Ying+ 2019]

GNNに対してそのまま拡張はできない

GNNのサロゲートモデルが自明でない

微分可能であれば使えるのでGNNにも有効

アーキテクチャ依存

GNNに特化

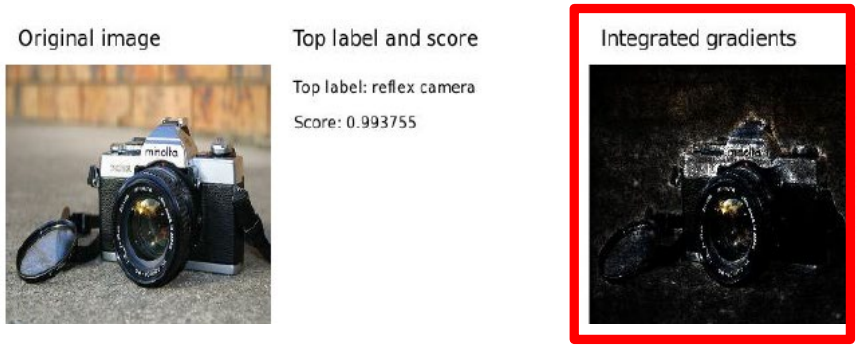
グラフに対して有効なものは多くはない

予測貢献度の可視化

Integrate gradients (IG) による可視化を実装

- ニューラルネットワークのアーキテクチャに依存しない

GCNを含むあらゆるネットワークに適用可能

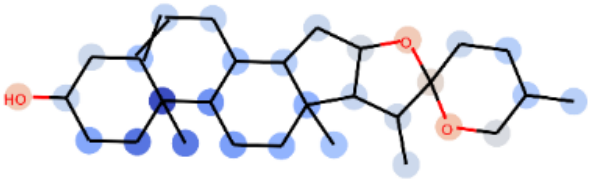


Sundararajan, Mukund, Ankur Taly, and Qiqi Yan. "Axiomatic attribution for deep networks." Proceedings of the 34th ICML-Volume 70. JMLR. org, 2017.

- グラフ上での予測の貢献度を算出可能

赤い部分が予測により効いている

グラフのノードやエッジについた属性の貢献度を算出可能



グラフ・系列・ベクトルに対する可視化に統一的な仕組みで可視化が可能

もくじ

1. 背景：医療リアルワールドデータ

2. テーブルデータに対する機械学習

3. 医療時系列データ

[研究紹介] 血圧変動予測

4. グラフデータに関する研究紹介

[研究紹介]ゲノム関連知識グラフの構築とその活用

[研究紹介]kGCN： ライフサイエンス向け

グラフニューラルネットワークツール

まとめ

ＡＩ技術はどんどん便利に高度になっている。
その中で今の目的に合わせて適切に活用していくことが重要である。

一方で、ＡＩ技術自体も発展途上にあるので
今の常識にとらわれずに、未来の医療に向けての応用に期待したい。