

位相的データ解析とその機械学習応用

富士通研究所

人工知能研究所 自律学習PJ

池祐一

講師自己紹介

■ 池 祐一(いけ ゆういち)

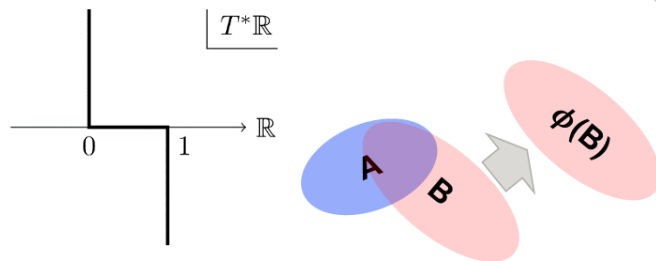
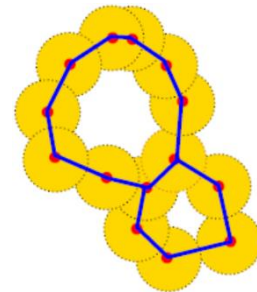


Figure 1: $SS(k_{[0,1]})$

- 2018年3月 東京大学大学院数理科学研究科博士課程修了
- 2018年4月 富士通グループ入社
位相的データ解析(TDA)の研究に従事
- 2019年10月～ JST ACT-X「数理・情報のフロンティア」
- 2020年8月～ 早稲田大学基幹理工学部客員次席研究員

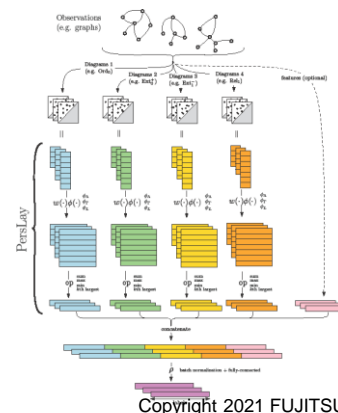
■ 第1部:TDAとその機械学習応用

- TDAはどう使われているか？
- TDAとは何か？
- TDAと機械学習とを組み合わせるには？



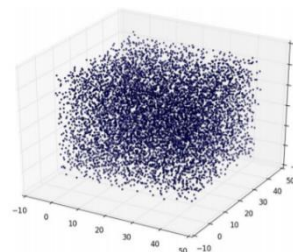
■ 第2部:TDAとその応用についてのチュートリアル

- TDAに関するOSSであるGUDHIを実際に動かしてみる
- GUDHIとscikit-learnを組み合わせさせて試してみる

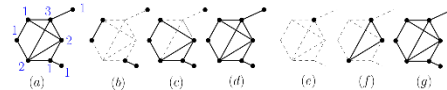


■ TDAは何に使えるか？

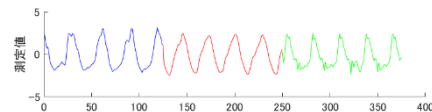
- **物質科学への応用: シリカやタンパク質の解析**
分子動力学法 (MD) シミュレーションと合わせて使う



- **グラフ分析への応用: グラフデータの解析**
ネットワークやケモインフォマティクスにおけるグラフデータなど



- **時系列解析への応用: ノイズを持つカオス的時系列の解析**
振動データ・心電図・脳波などに適用可能



■ TDAとは何か？何が利点かを説明します

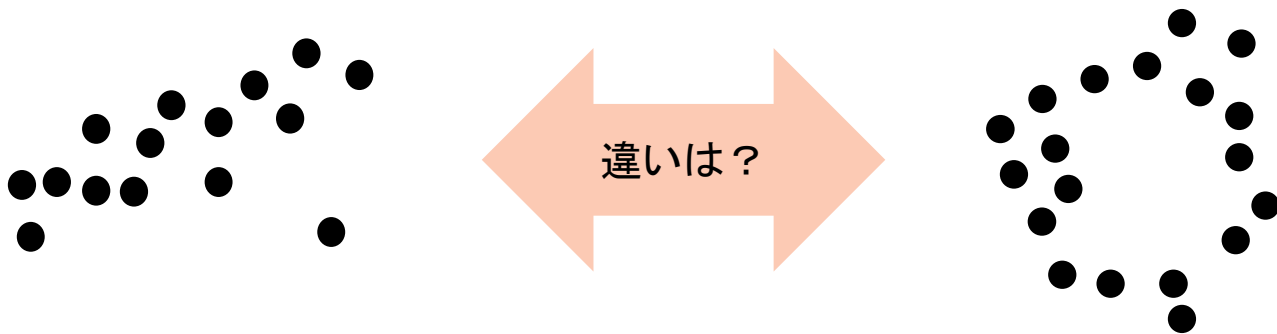


TDAとは何か？

- データの「形」とは？
- TDAのアイデア
- パーシステントホモロジー・パーシステンス図

■ 位相的データ解析 (Topological Data Analysis, TDA)

- データの幾何的な特徴を抽出する比較的新しい手法 (Edelsbrunner et. al. '02, Zomorodian and Carlsson '05)
- データの大まかな「形」(連結成分・穴の数)に着目



■ 背景

1. 複雑な幾何的構造を持つデータを扱う必要性
2. 計算幾何学の発展

データの「形」とは？

■ データの「形」を考えるためにトポロジーの考え方を使う

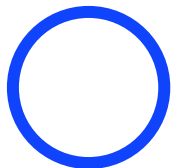
■ トポロジー: 空間の「連続的な形」を扱う数学の分野



■ どうやって「連続的な形」を区別するか？

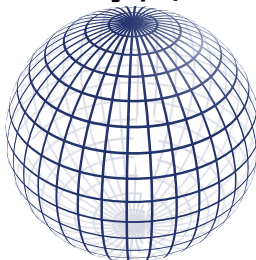
-> ホモロジー: 各次元の「穴」を抽出する不変量(穴の数: ベッチ数)

円周



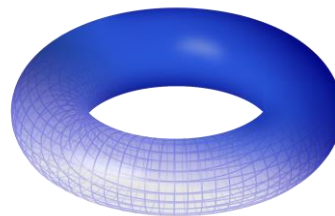
0-dim (#連結成分) : 1
1-dim (#ループ) : 1
2-dim (#空洞) : 0

球面



0-dim : 1
1-dim : 0
2-dim : 1

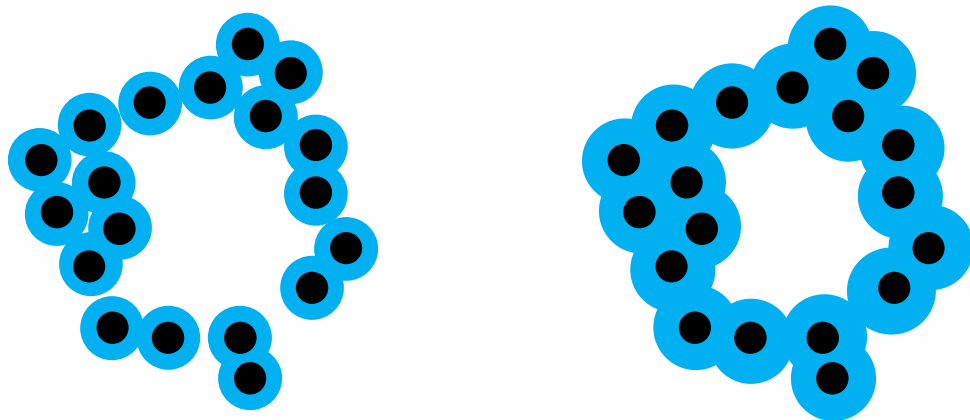
トーラス



0-dim : 1
1-dim : 2
2-dim : 1

データ点群の位相的特徴量

- Q. バラバラな点群からどうやってトポロジーを取り出すか？
- アイデア1: データ点中心の球の和集合を考える

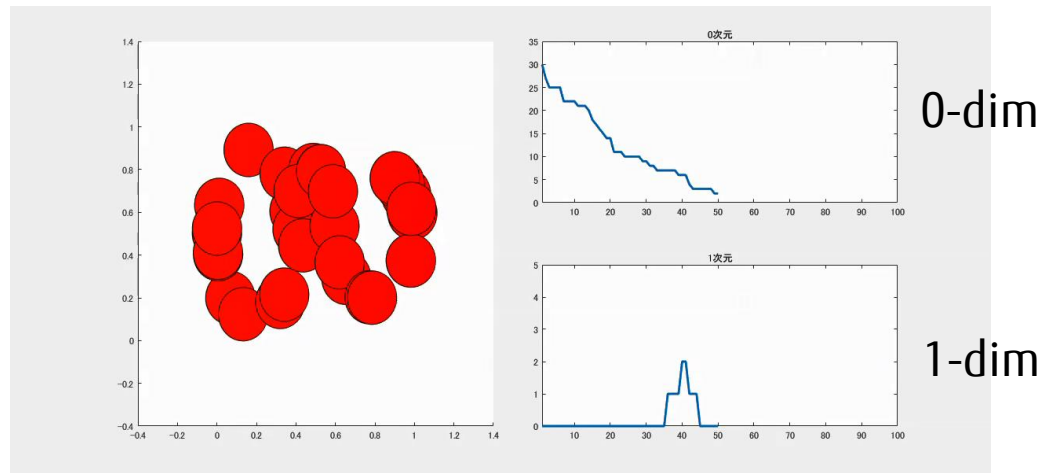


- 問題点: 半径をどのようにとればよいか事前に分からない
 - 小さすぎ -> バラバラのまま, 大きすぎ -> 全部がくっついて穴が見えない

パーシステントホモロジーのアイデア

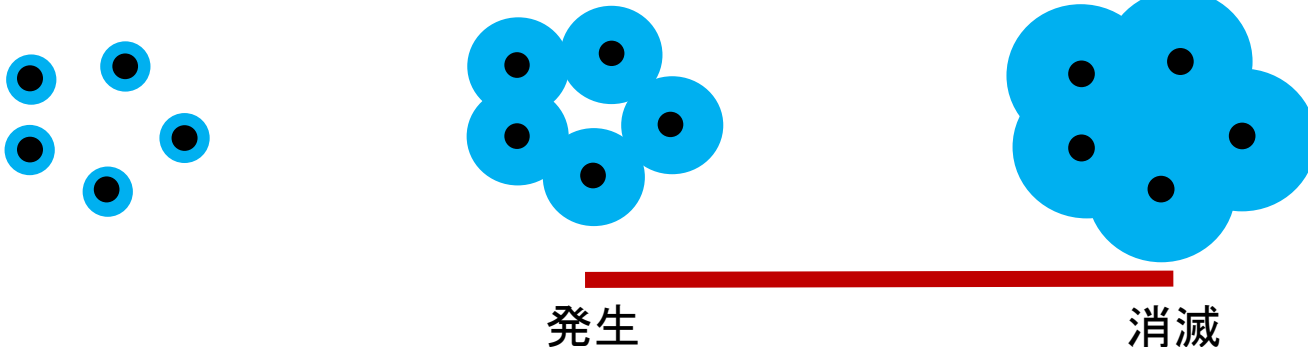
- Q. バラバラな点群からどうやってトポロジーを取り出すか？
- アイデア1: データ点中心の球の和集合を考える
- アイデア2: 半径を一つに止めずに動かす
 - 半径が大きくなっていく際のトポロジーの変化を追跡

- 長く続く特徴は「本質的」とみなせる
- マルチスケールの解析
- 穴の数の変化のグラフ:
ベッチ曲線



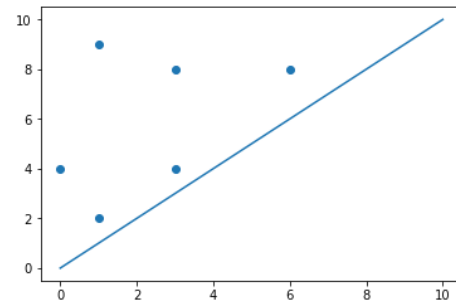
パーシステンス図 (PD)

- 各トポロジ的特徴量(連結成分・穴など)に対して, 発生・消滅時刻を見つけることができる



■ パーシステンス図

- 発生時刻を第1軸, 消滅時刻を第2軸にプロット
- パーシステントホモロジーの情報を持つ
- 対角線の遠くが本質的(対角線の近くはノイズ)



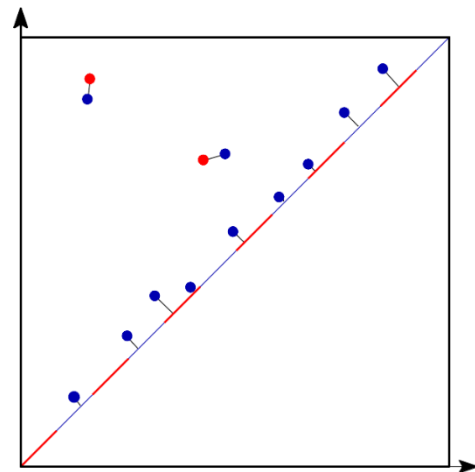
パーシステン스図間のボトルネック距離

- パーシステン스図がどれくらい異なっているかを距離として測る
- 主要な距離の一つ: **ボトルネック距離**
 - 各点をマッチングさせてそれらの距離の最大を考える
 - 対角線に近い点はノイズとみなすので対角線への射影とマッチングさせる

$$\blacksquare d_B(D_1, D_2) := \inf_{\gamma} \sup_{x \in D_1} \|x - \gamma(x)\|_{\infty}$$

$\gamma: D_1 \rightarrow D_2 \cup \Delta$: マッチング

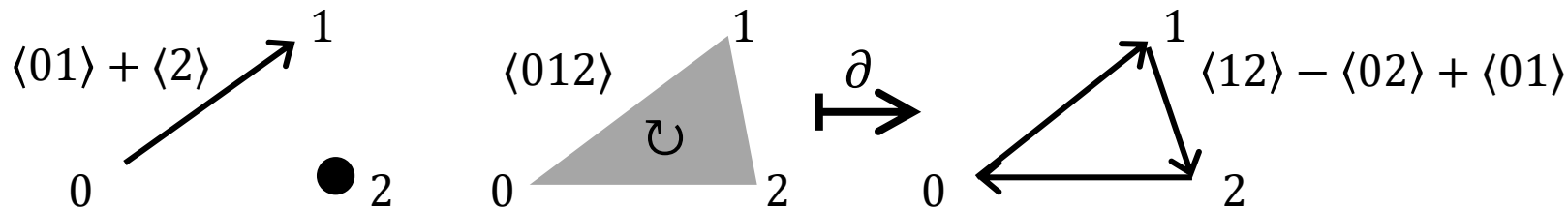
- **安定性定理**: 入力点群の距離でPD間の距離を評価→**ノイズ耐性**



計算機にはどのように実装されるか？

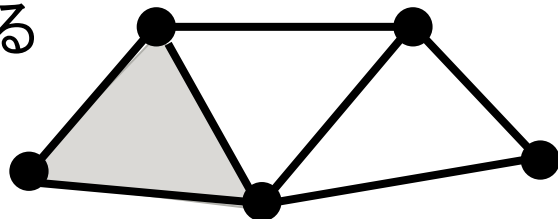
■ 「穴」の数(ホモロジー)を組合せ的に扱う方法: **単体複体**

- 単体(三角形の一般化)が集まってできたもの, グラフの一般化
- 単体の集まりの間の境界準同形で「穴」の数を計算



- 「穴」: **境界が消えるもので中が埋まっていない**(何かの境界になっていない) **もの**

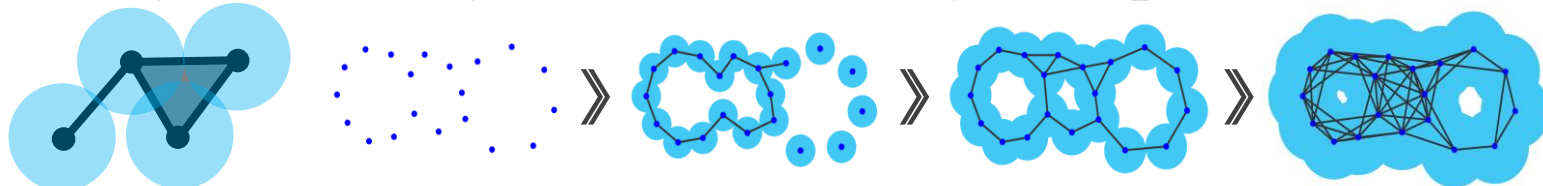
■ 単体複体を使って組合せ的に「穴」を計算できる



データから単体複体を作る方法

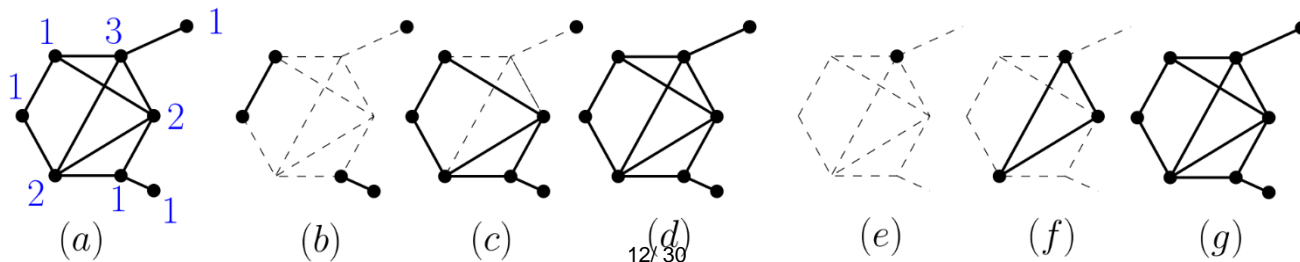
■ 点群から単体複体を作る主な方法: **Vietoris-Rips複体**

1. $\varepsilon > 0$ に対して二つの頂点を中心とする半径 ε の球が交わったら辺を引く
2. 三つの頂点についてもすべての組合せに辺があれば中を埋める



■ ε を動かすと**フィルトレーション**(単体複体の部分単体の増大列)ができる

■ **重み付きグラフ**からもフィルトレーションを作れる

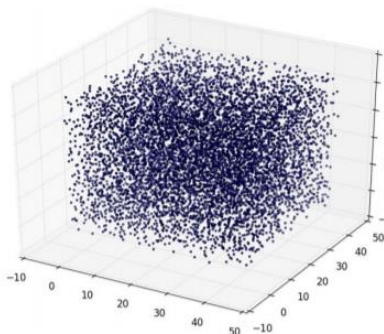


TDAの実応用

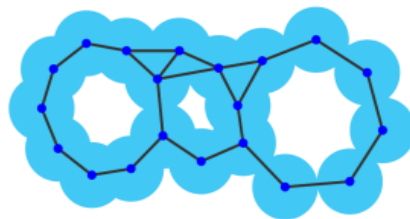
- 物質科学への応用
- パーシステンス図のベクトル化
- グラフデータ解析
- 時系列解析

TDAの基本的な使い方

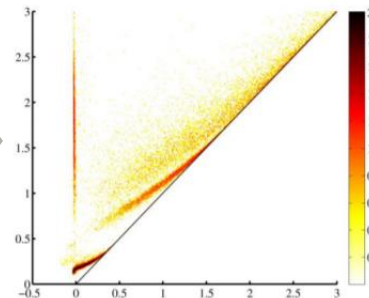
データ



フィルトレーション



パーシステンス図 (PD)



専門家

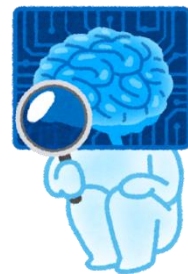


例

- シミュレーションで生成された点群
- センサーによる振動
- グラフ・画像

Software
Ripser, GUDHI,
HomCloud, ...

機械学習に入力するために
ベクトル化が必要



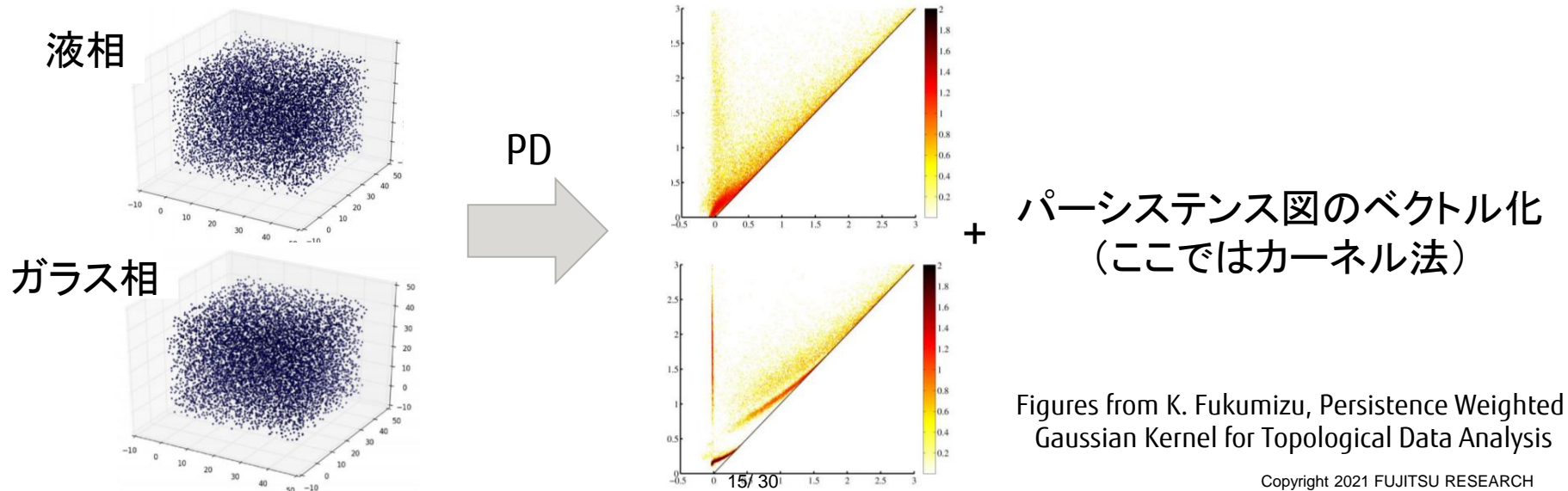
機械学習

Figures from K. Fukumizu, Persistence Weighted
Gaussian Kernel for Topological Data Analysis

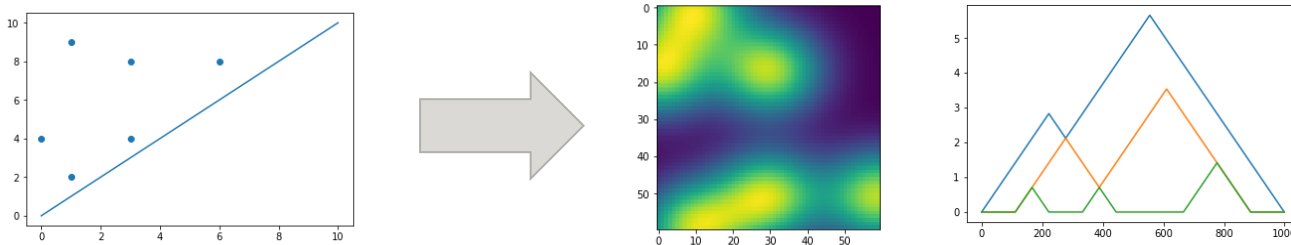
■ G. Kusano, K. Fukumizu, Y. Hiraoka: Persistence weighted Gaussian kernel for topological data analysis, ICML2016

■ SiO_2 が液体からガラスに変化する温度を推定したい

■ アイデア: **点群をPDに変換**して, それが変化する温度を調べる

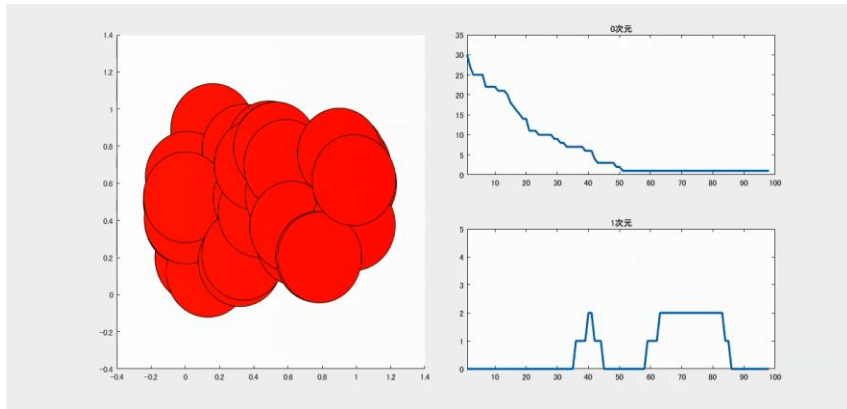


■ 様々なPDのベクトル化方法が提案されている

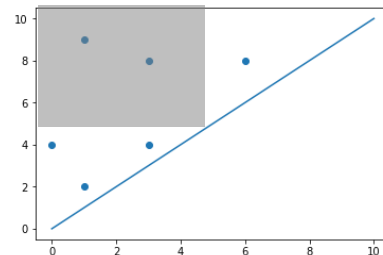


- ベッチ曲線: 穴の数の変化の曲線
- パーシステンシメージ: PDを画像に変換
- パーシステンスランドスケープ: PDを対角線上の関数の列に変換
- カーネル法, . . .

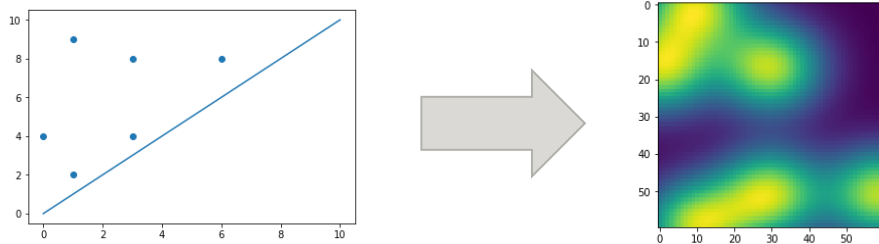
■ 球の半径に対して穴がいくつあるかをあらわす量



- パーシステンス図からは $b \leq t < d$ を満たす点 (b, d) の個数を対応させればよい
- 実用上はどの t まで・どの解像度で取るかを指定

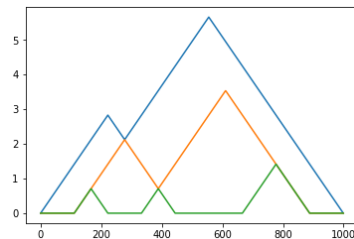
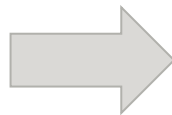
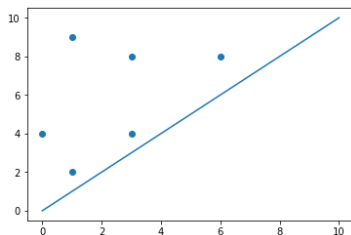


■ パーシステンス図を画像に変換



- $(b, d) \mapsto (b, d - b)$ の変換をかける
 - 各点を中心とするガウス関数の和であらわされる関数を出力とする
 - 実用上はどの範囲で考えるか・どの解像度にするかを指定
- このあとにCNNと組み合わせることも可能

■ 実数上の関数の列に変換



■ 2次元空間の点 (b, d) に対して,

$$f_{(b,d)}(t) = \max(0, \min(b + t, d - t))$$

と定める(対角線に三角を下ろす)

■ 自然数 $k \in \mathbb{N}$ に対して, $\lambda_D(k, t) = k \max_{(b,d) \in D} f_{(b,d)}(t)$ と定める

■ パーシステンスランドスケープは統計的に良い性質を持つ (Bubenik)

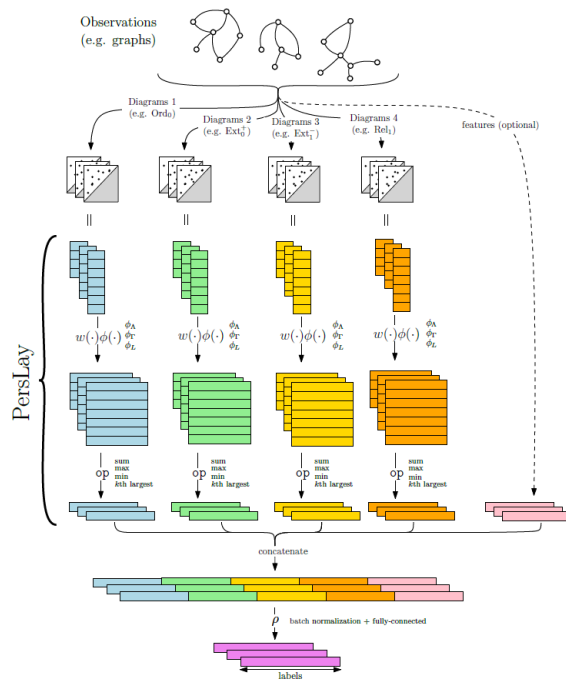
(参考)ベクトル化の学習・グラフ分類への応用

■ PDのベクトル化そのものを学習させる取り組みもある
(Hofer et al. 2017, Carrière et al. 2020)

■ PersLay: ベクトル化をNNの一層として学習

■ グラフからその上の関数を構成してTDA特徴量で分類

Dataset	SV ¹	RetGK* ²	FGSD ³	GCNN ⁴	GIN ⁵	PERSLAY	
						Mean	Max
REDDIT5K	—	56.1	47.8	52.9	57.0	55.6	56.5
REDDIT12K	—	48.7	—	46.6	—	47.7	49.1
COLLAB	—	81.0	80.0	79.6	80.1	76.4	78.0
IMDB-B	72.9	71.9	73.6	73.1	74.3	71.2	72.6
IMDB-M	50.3	47.7	52.4	50.3	52.1	48.8	52.2
COX2*	78.4	80.1	—	—	—	80.9	81.6
DHFR*	78.4	81.5	—	—	—	80.3	80.9
MUTAG*	88.3	90.3	92.1	86.7	89.0	89.8	91.5
PROTEINS*	72.6	75.8	73.4	76.3	75.9	74.8	75.9
NCI1*	71.6	84.5	79.8	78.4	82.7	73.5	74.0
NCI109*	70.5	—	78.8	—	—	69.5	70.1



- Ripser: 高速, 機能はまあまあ
- HomCloud: 日本の研究者によって開発, 逆問題に強い
- **GUDHI**: フランス研究機関Inriaが主導, 講演者も関わっている

गुडी GUDHI Geometry Understanding in Higher Dimensions

NEW ARTICLE

Fujitsu and France's Inria Jointly Develop Technology

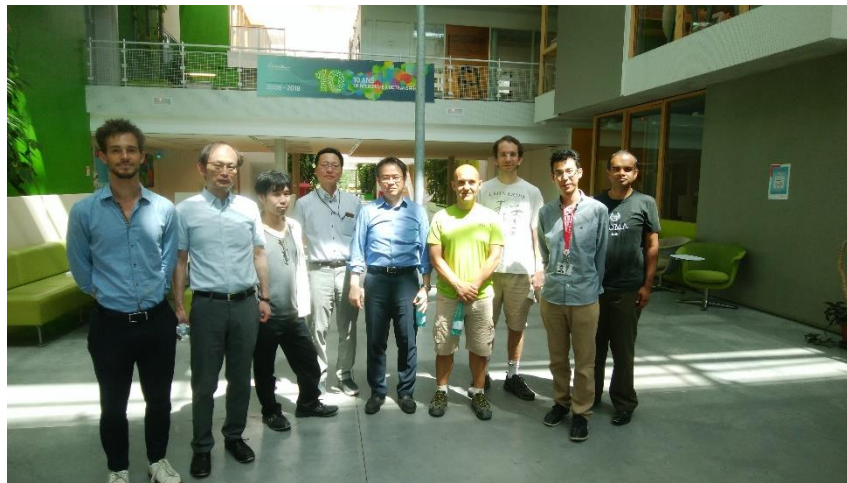
Anomaly-detecting AI models automatic creation

Resulting of a collaboration with Fujitsu, we are happy to announce a new time delay python module to help with the time series analysis. This feature will be available in the next GUDHI release.

More information is available in [this article](#).

📅 2020-03-16 📄 ARTICLE

🤝 Collaboration



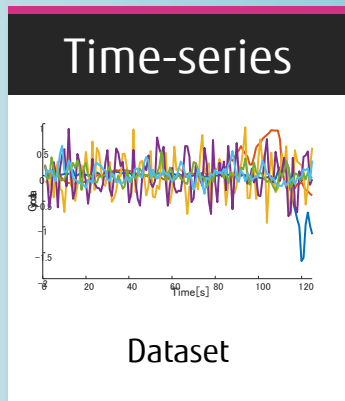
TDAの時系列への応用

■ TDAを時系列解析に応用する

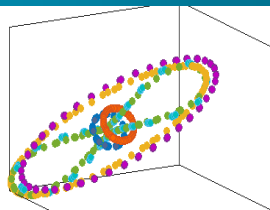
- アイデア: 時系列を点群に変換して, そのPDを調べる

TDA

2 steps analysis
for time series



Time-delay emb.



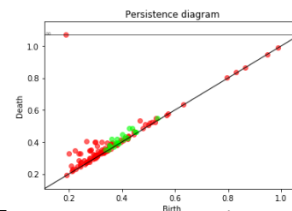
1

Attractor translation

Figuring time series rules



TDA feature



2

Feature extraction

Vectorization of persistence diagrams

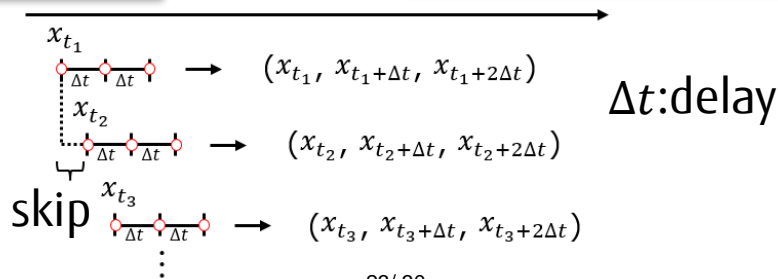
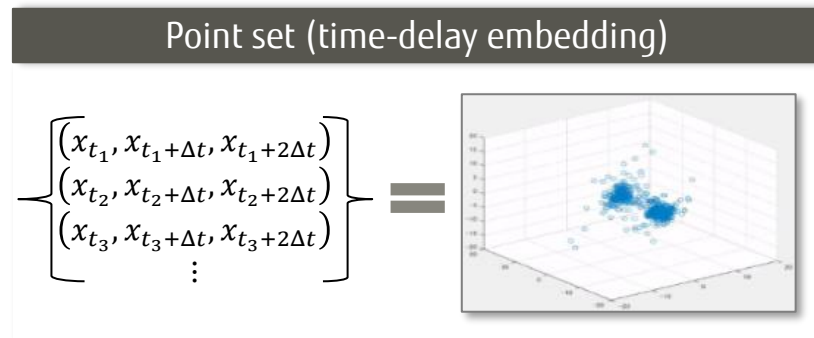
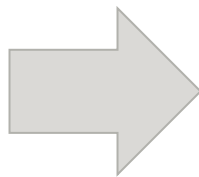
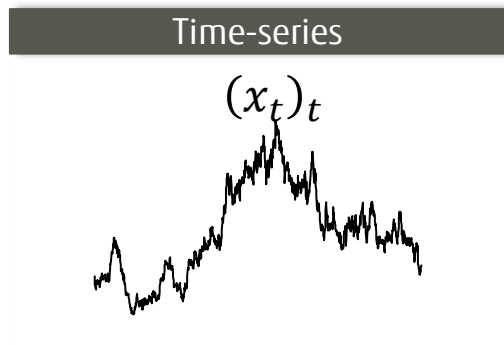
■ TDA的な手法は特にカオス的な時系列データに有効

時間遅れ埋め込み: 時系列を点群に変換する手法

■ 時系列をその力学系を反映する点群に変換してTDAを適用

■ カオス的なデータに有効

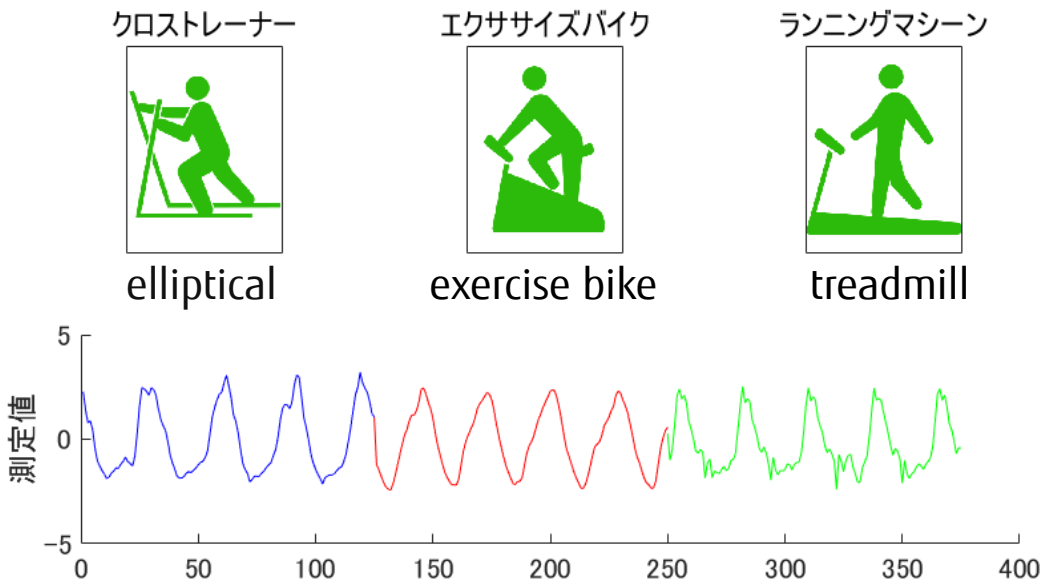
■ 金融データ・心電図・脳波の解析などに応用されている



数学的には**Takensの埋め込み定理**に基づく

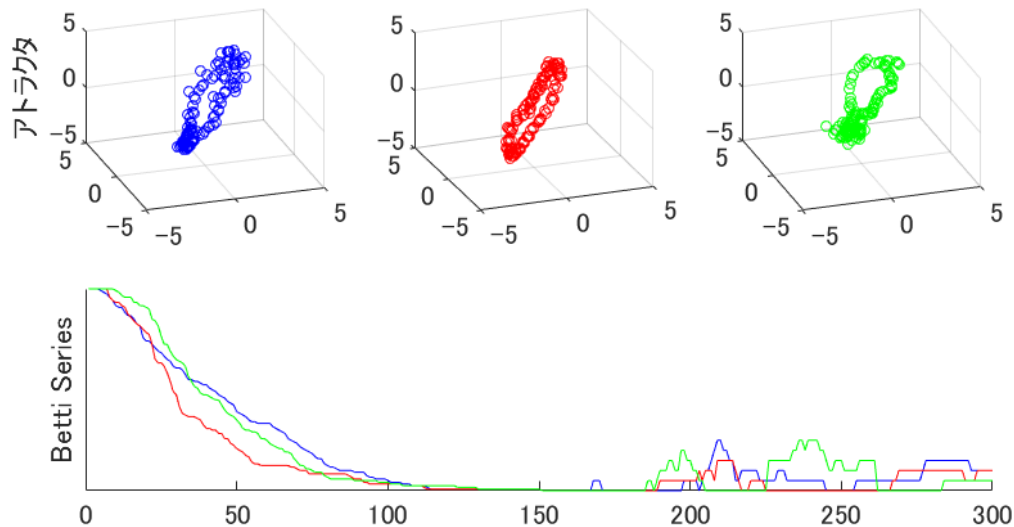
■ ジャイロスコープのデータから行動を分類

■ もとの信号では区別が困難



■ 時間遅れ埋め込みを用いて点群に変換

- 点群の形は差が明確 -> TDAで分類が可能
- 下ではベッチ曲線でTDA特徴量を取り出した



心電図解析への応用

■ 富士通が開発したTDA手法は不整脈検知において良い精度を達成



問題

不整脈はタイプにより
生死にかかわる


早期発見が必要



Abnormal

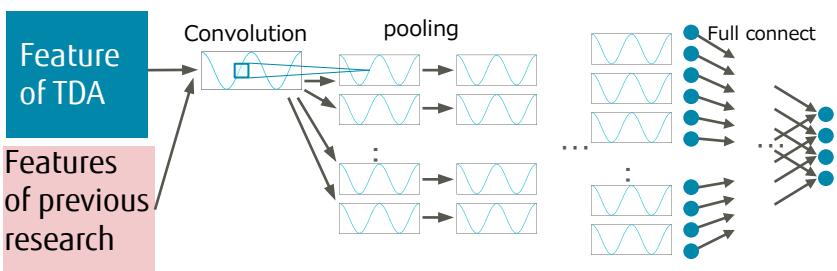



Normal



技術

従来技術の特徴量とTDA特徴量を
組み合わせてCNNを学習





Inriaとの共同研究

- 新たなネットワーク構造
- 新たな特徴量生成手法 (DTM-filtration)



効果

誤分類が**70%削減**

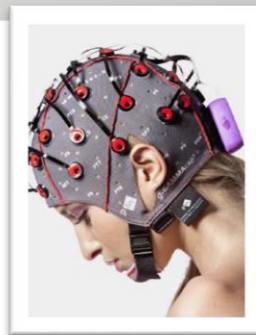




病気の早期発見に役立つ

脳波を用いたせん妄検出

- せん妄は一般的で危険であるにも関わらず判定が困難
- TDA手法でせん妄診断の精度を向上させることができる



背景

高齢者は入院中のせん妄リスクが高い

早期発見が必要

転倒による入院期間の延長を防ぎたい

技術

一点から取った1チャンネル脳波データのTDA特徴量を用いてせん妄を判定する

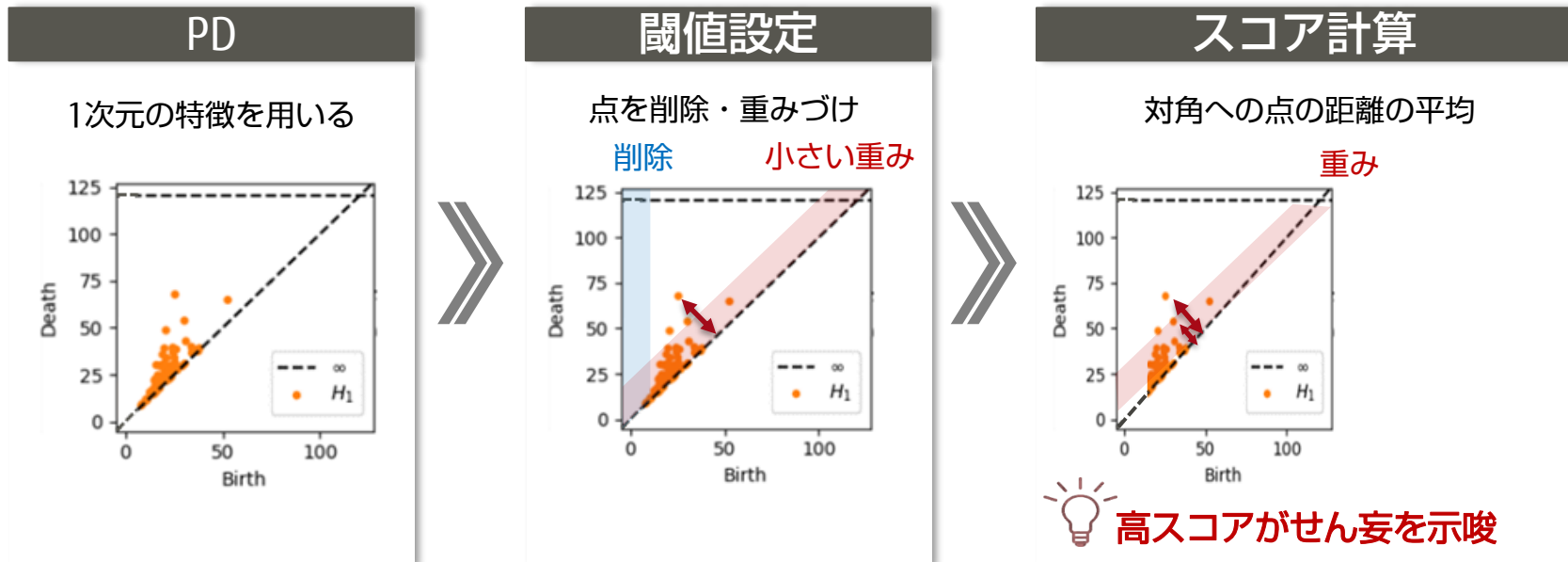
判定

効果

見落とし率が**半減**

TDAせん妄検知スコア:手法

- 2秒の信号x30 [=60sec] を時間遅れ埋め込みによって点群に変換してPDを作り, そこから下記の方法でスコアを構成



TDAせん妄検知スコア: 評価

■ データセット: アイオワ大学病院の入院患者273名

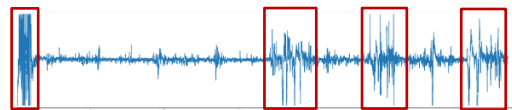
Number of patients

Delirium : 100
Non-delirium : 173

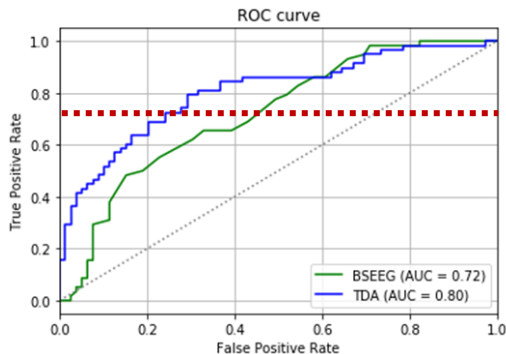
After preprocessing

Delirium : 64
Non-delirium : 73

ノイズ部分を削除
電源ノイズを除去



□の部分を削除



AUC

0.721

UP!

0.805

BSEEG

TDA

specificity w/
sensitivity=0.75

Specificity

0.552

UP!!!

0.712

BSEEG

TDA

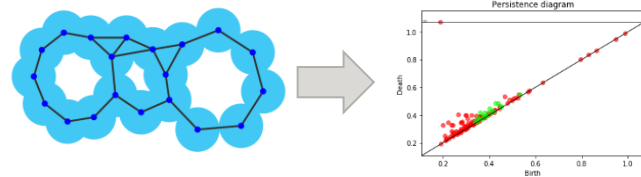
AUCとspecificityをTDA手法で改善

まとめ

■ 位相的データ解析 (TDA)

- 主な道具: パーシステントホモロジー

- データの「形」の情報をパーシステンス図として抽出する



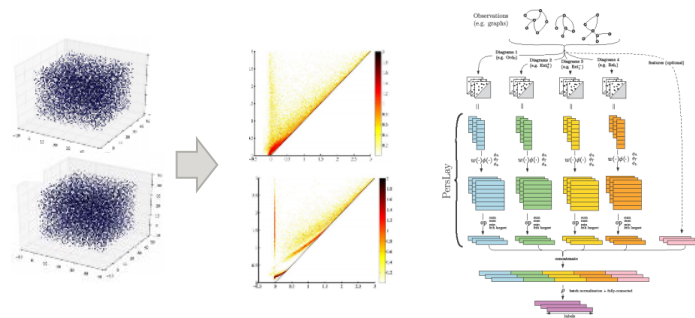
■ 応用

- 物質科学

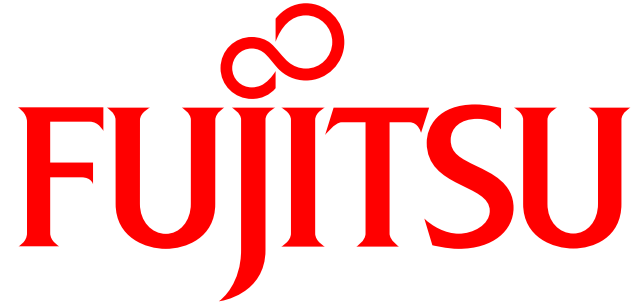
- グラフデータ解析

- 時間遅れ埋め込みと組み合わせた時系列解析

- 機械学習との組合せにより柔軟な活用が可能



■ 第2部のチュートリアルではGUDHIを使って実際にTDAを動かしてみます



shaping tomorrow with you