# パーシステントホモロジーとその応用 (1)

2024-03-23

純粋数学者のためのデータサイエンス入門

九州大学 マス・フォア・インダストリ研究所

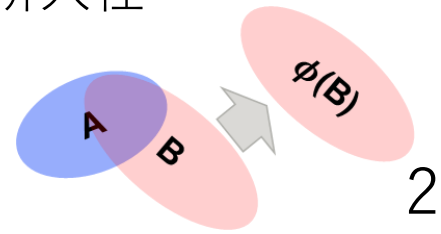
池 祐一

# 自己紹介

## 池 祐一（いけ ゆういち）1990年新潟生まれ

■東大数理で博士取得
（層理論とシンプレクティック幾何）
■ エネルギー評価に**TDA的距離**が使える！

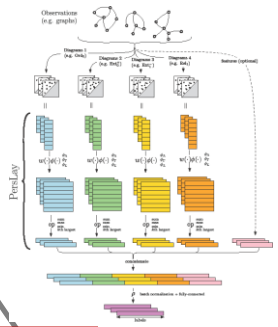
■富士通研究所入社




2018

2020

■東京大学に異動



2022

■九州大学に異動
■教科書を出版

2024

富士通

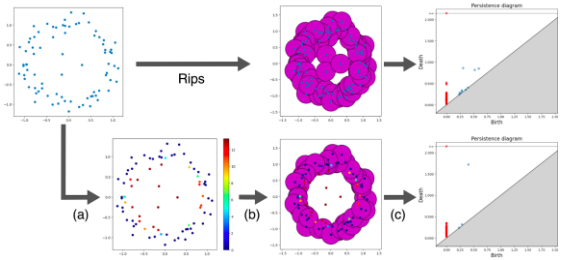
東京大学

九大IMI

2019

■フランスInriaとの共同研究のため2ヶ月フランス滞在
■ACT-X採択

2021

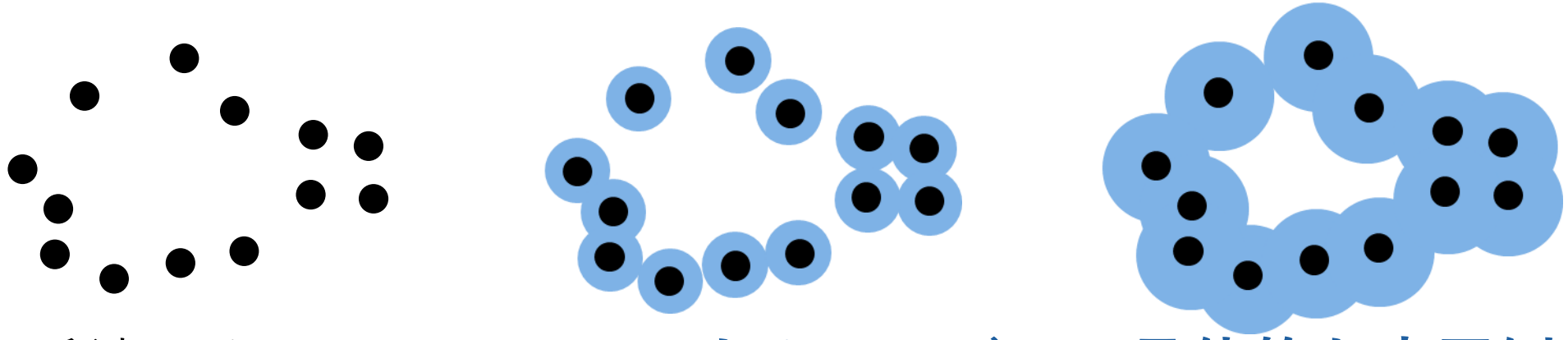



2023

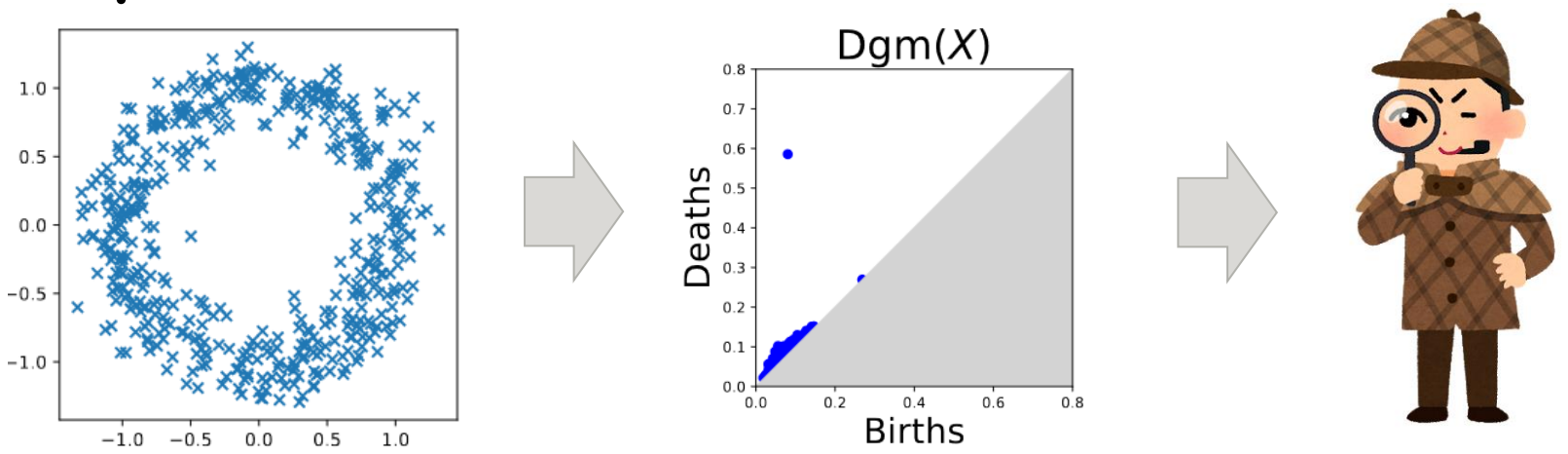
**研究分野**
- **層理論**
- **位相的データ解析**
- **幾何学**

# 概要

■**位相的データ解析** (Topological Data Analysis, TDA) とは，データの「トポロジー」に着目して解析を行う手法.



■主要な手法である**パーシステントホモロジー**と**具体的な応用例**を説明する.

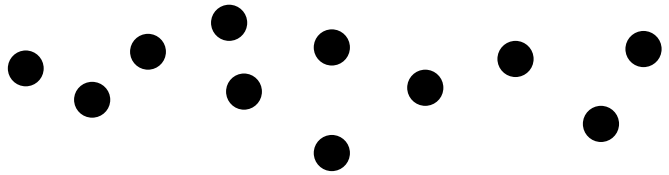


# 目次

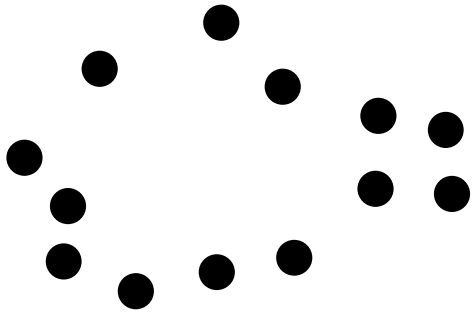
# 位相的データ解析の考え方とパーシステントホモロジー

データの「トポロジー」を抽出するには，パーシステントホモロジーの定義

# データのトポロジー：TDAのアイデア



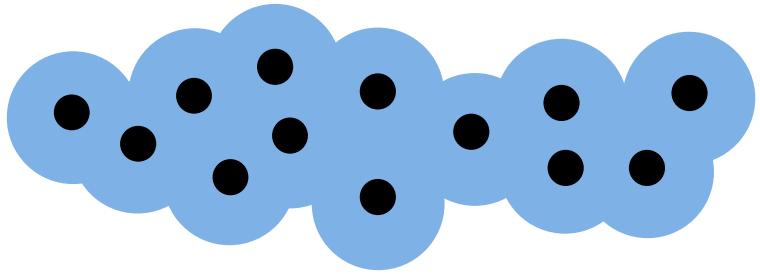
穴がない



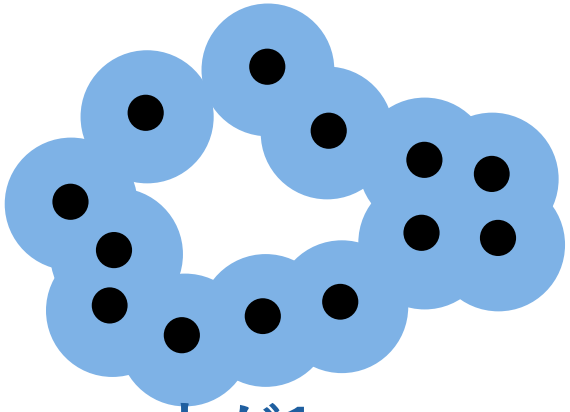
穴が1つ

Q. どうやって離散的なデータから「トポロジー」を取り出すか？

# データのトポロジー：TDAのアイデア



穴がない



穴が1つ

Q. どうやって離散的なデータから「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：$\cup_{p \in P} \overline{B(p;r)}$
→**ホモロジー**が計算できればトポロジーの差を検出可能な場合がある

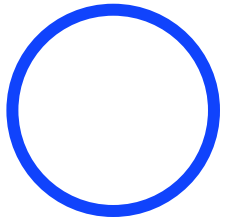
# 位相空間のホモロジー

## ホモロジー $H_n(X)$

■位相空間$X$の**$n$次元の「穴」**をあらわすベクトル空間
（今回は簡単のために$\mathbb{F}_2$で考える）

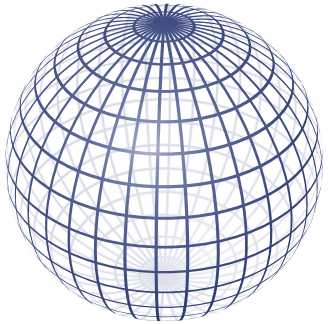
■同相・ホモトピー同値ならば同型になる（**位相不変量**）
→ホモロジーが異なればホモトピー同値で**ない**：**形が区別可能**

円周



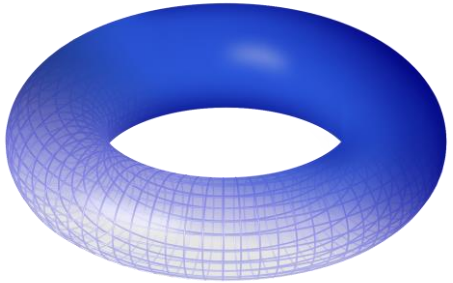
$H_0(S^1) = \mathbb{F}_2$（連結成分）
$H_1(S^1) = \mathbb{F}_2$（ループ）
$H_2(S^1) = 0$　（空洞）

球面



$H_0(S^2) = \mathbb{F}_2$
$H_1(S^2) = 0$
$H_2(S^2) = \mathbb{F}_2$

トーラス



$H_0(T^2) = \mathbb{F}_2$
$H_1(T^2) = (\mathbb{F}_2)^2$
$H_2(T^2) = \mathbb{F}_2$

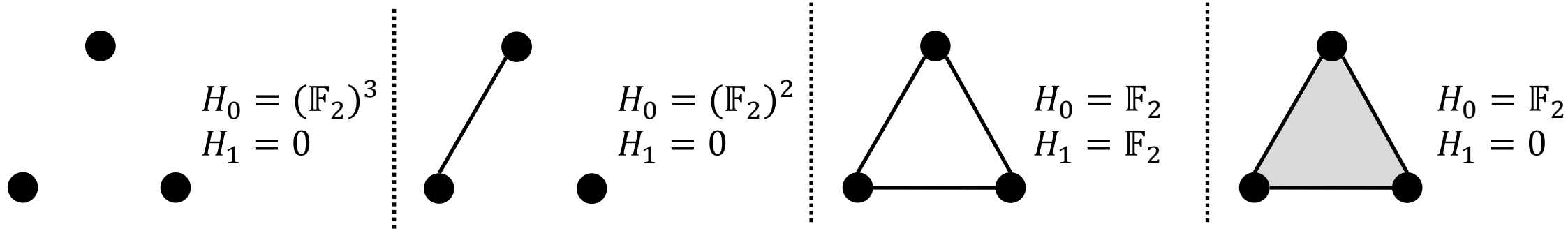
# 単体的ホモロジー

**コンピュータでの計算**には単体的複体のホモロジーが向いている

■（有限）**単体的複体**とは空集合を含まない有限集合$K$であって
$\sigma \in K, \emptyset \neq \tau \subset \sigma \Rightarrow \tau \in K$を満たすもの

■ここから組合せ的に**単体的ホモロジー**$H_n(K)$が定まる

■$H_n(K) \cong H_n(|K|)$：幾何学的実現のホモロジーと同型




$$K_n := \{\sigma \in K \mid \#\sigma = n+1\}, \qquad C_n(K) := \bigoplus_{\sigma \in K_n} \mathbb{F}_2 \sigma,$$

$$\partial_n : C_n(K) \to C_{n-1}(K); \ \{v_0, \dots v_n\} \mapsto \sum_{i=0}^{n} \{v_0, \dots \hat{v_i}, \dots v_n\},$$

$$H_n(K) := \operatorname{Ker} \partial_n / \operatorname{Im} \partial_{n+1} \ (\partial_0 = 0)$$

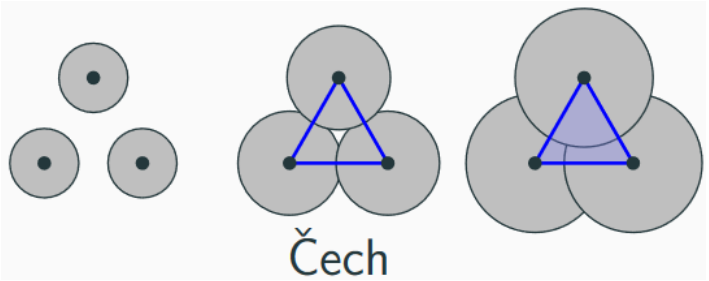
# 点群から単体的複体を作る

点群$P \subset \mathbb{R}^d$と$r \in \mathbb{R}$から単体的複体を構成できる

■$\{x_0, \ldots x_n\} \subset P$が単体になることを次のように定義

**Čech複体**

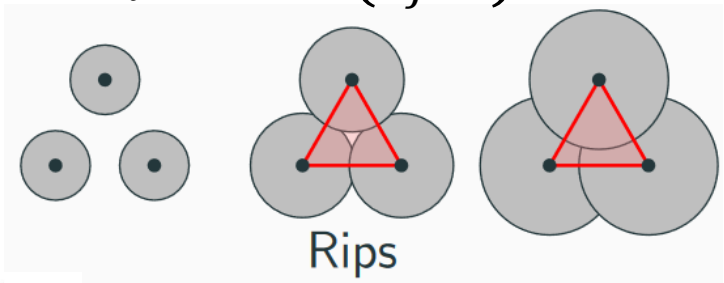
$\{x_0, \ldots, x_n\} \in C(P; r) :\Leftrightarrow \bigcap_i \overline{B(x_i; r)} \neq \emptyset$



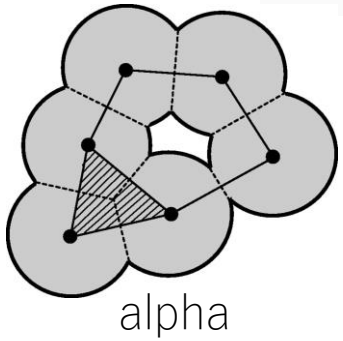

**Vietoris-Rips複体**

$\{x_0, \ldots, x_n\} \in R(P; r)$

$:\Leftrightarrow \overline{B(x_i; r)} \cap \overline{B(x_j; r)} \neq \emptyset \ (\forall i, j)$

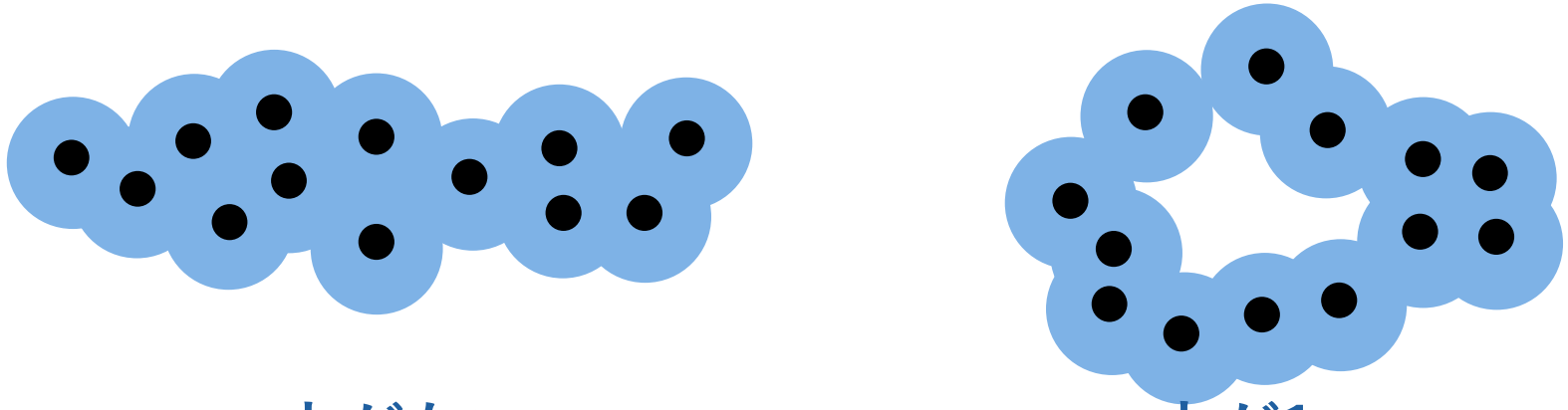



**アルファ複体**




■$|C(P; r)|$や$|\alpha(P; r)|$は球の和集合$\bigcup_{p \in P} \overline{B(p; r)}$とホモトピー同値

# データのトポロジー：TDAのアイデア

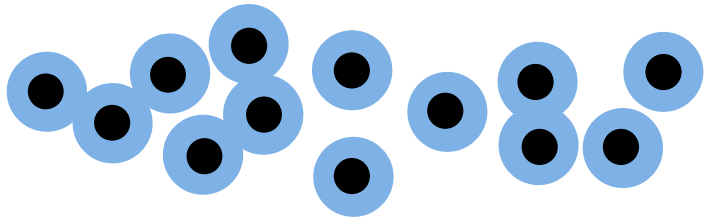


穴がない　　穴が1つ

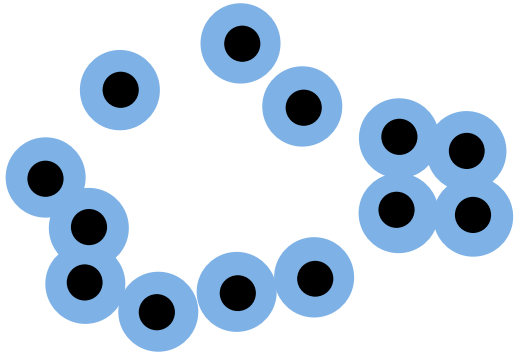
Q. どうやって離散的なデータから「トポロジー」を取り出すか？

■アイデア1：**データ点中心の球の和集合**を考える：$\cup_{p \in P} \overline{B(p;r)}$
→**ホモロジー**が計算できればトポロジーの差を検出可能な場合がある

# データのトポロジー：TDAのアイデア



穴がない



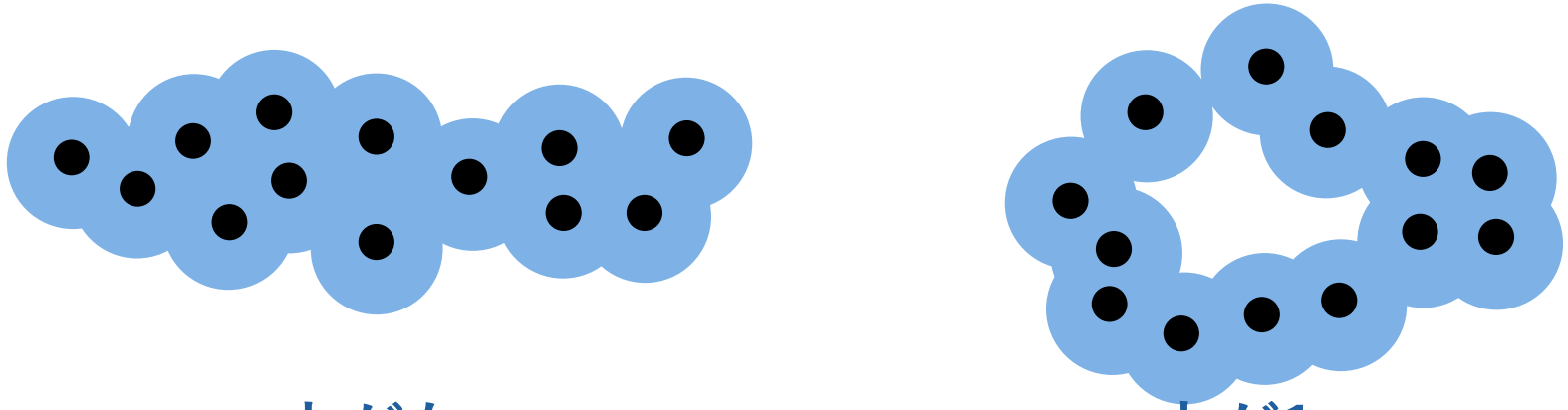
穴が1つ

Q. どうやって離散的なデータから「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：$\cup_{p \in P} \overline{B(p;r)}$
→**ホモロジー**が計算できればトポロジーの差を検出可能な場合がある

**問題点**：**どのように半径$r$を設定すればよいか分からない**

# データのトポロジー：TDAのアイデア



穴がない　　穴が1つ

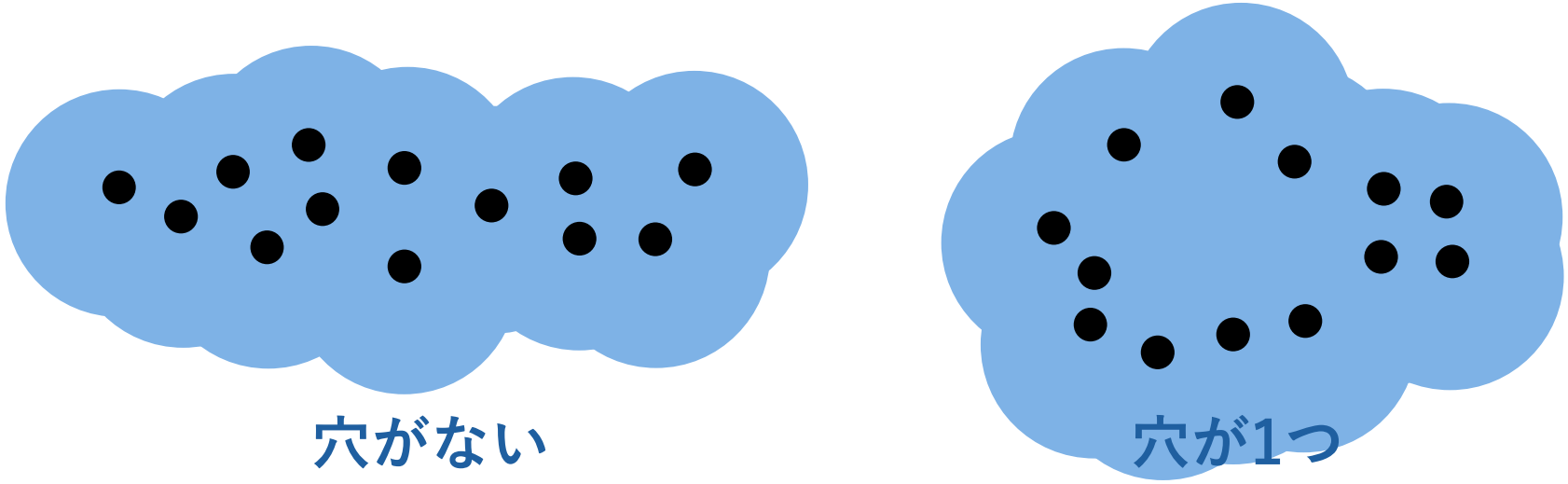
Q. どうやって離散的なデータから「トポロジー」を取り出すか？

■アイデア1：**データ点中心の球の和集合**を考える：$\cup_{p\in P}\overline{B(p;r)}$
→**ホモロジー**が計算できればトポロジーの差を検出可能な場合がある

**問題点**：**どのように半径$r$を設定すればよいか分からない**

■アイデア2：**半径を大きくしていきトポロジーの変化を追跡**する

# データのトポロジー：TDAのアイデア



穴がない　　穴が1つ

Q. どうやって離散的なデータから「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：$\cup_{p \in P} \overline{B(p;r)}$
→**ホモロジー**が計算できればトポロジーの差を検出可能な場合がある

**問題点**：**どのように半径$r$を設定すればよいか分からない**

■アイデア2：**半径を大きくしていきトポロジーの変化を追跡**する
⇒ **長く持続（persist）するホモロジー類は本質的**とみなす

# パーシステントホモロジー

■数学的には，$r \leq s$に対して包含写像$\bigcup_p \overline{B(p;r)} \hookrightarrow \bigcup_p \overline{B(p;s)}$から誘導されたホモロジー間の写像の族

$$H_n\left(\bigcup_{p\in P}\overline{B(p;r)}\right) \to H_n\left(\bigcup_{p\in P}\overline{B(p;s)}\right)$$

を考えるということ

■より一般に $f: X \to \mathbb{R}$ 関数 $\rightsquigarrow X(f)_r = \{x \in X \mid f(x) \leq r\}$ 劣位集合

ホモロジーと誘導写像の族（$(\mathbb{R}, \leq) \to \mathrm{Vect}(\mathbb{F}_2)$なる関手）

$$PH_n(f): \cdots \to H_n(X(f)_r) \to H_n(X(f)_s) \to \cdots$$

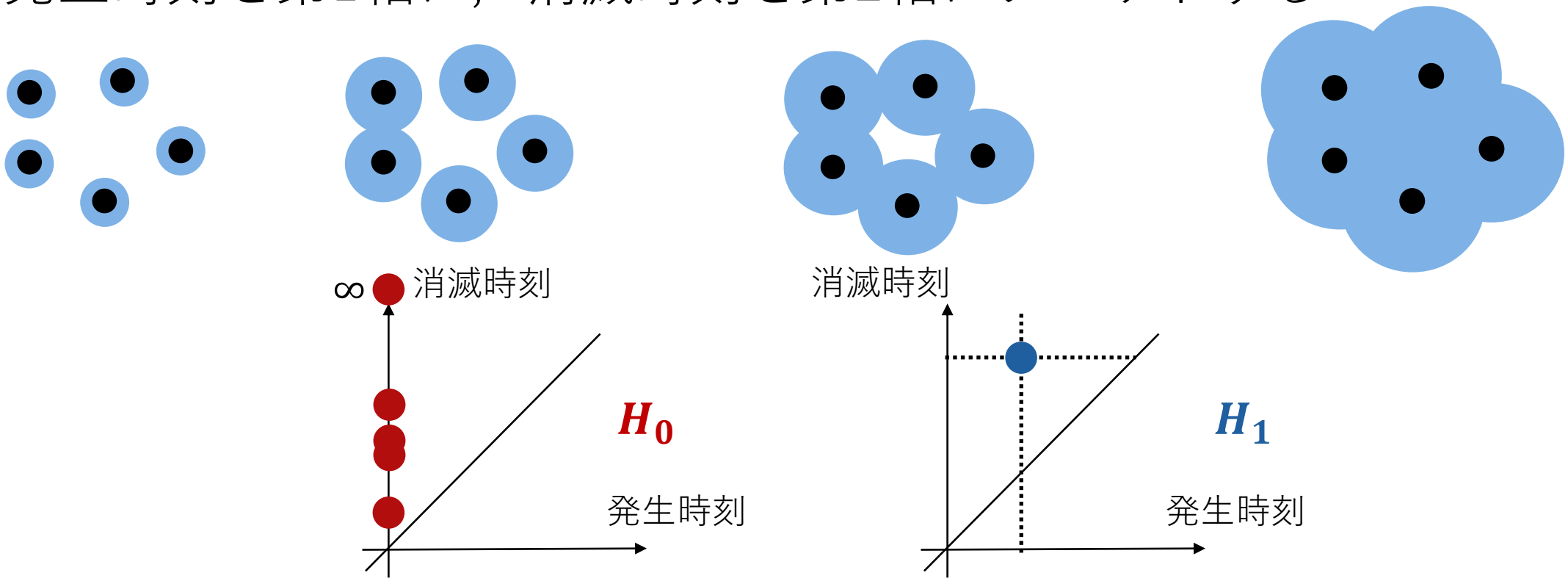
**$f$の劣位集合フィルトレーションのパーシステントホモロジー (PH)**

**例**： $P \subset \mathbb{R}^d = X$ 有限集合，$f = d(\cdot, P) \Rightarrow X(f)_r = \bigcup_{p\in P}\overline{B(p;r)}$

# PHの情報の可視化：パーシステンス図

各ホモロジー類に対し，生成する時刻$b$と消滅する時刻$d$が存在

↝発生時刻を第1軸に，消滅時刻を第2軸にプロットする




こうしてプロットされた$\overline{\mathbb{R}}^2$ $(\overline{\mathbb{R}} = \mathbb{R} \cup \{\pm\infty\})$の多重部分集合を

**パーシステンス図（persistence diagram, PD）**という

# パーシステンス図：より正確には

[Crawley-Boevey, 2015] 任意の$r$に対して$H_n(X(f)_r)$が有限次元なら

$$PH_n(f) \cong \bigoplus_{\alpha \in A} (\mathbb{F}_2)_{I_\alpha}, \qquad (I_\alpha \text{は空でない区間})$$

と直和分解．ここで区間$I$に対して$(\mathbb{F}_2)_I \colon (\mathbb{R}, \leq) \to \mathbf{Vect}(\mathbb{F}_2)$は，
$I$上でベクトル空間としては$\mathbb{F}_2$で写像が恒等写像であり，
そのほかで0となるもの．この分解は**一意的**

※もしホモロジーの変化が有限回なら
単因子論的な議論で証明できる

このとき，$I_\alpha$の左端を$b_\alpha$・右端を$d_\alpha$として，$\overline{\mathbb{R}}^2$の多重部分集合

$$D_n(f) \coloneqq \{ (b_\alpha, d_\alpha) \mid \alpha \in A \} \subset \overline{\mathbb{R}}^2$$

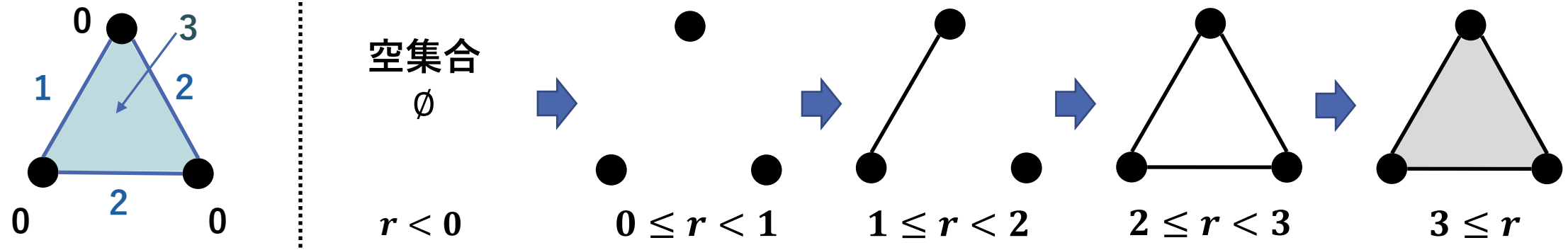
を$f$の（劣位集合フィルトレーションの）**パーシステンス図**と呼ぶ

# 単体的複体のフィルトレーションとPH

単体的複体のバージョンでPHを考える

**定義** 単体的複体$K$の**フィルトレーション**とは関数 $f\colon K \to \mathbb{R}$ であって $\sigma \subset \tau \Rightarrow f(\sigma) \leq f(\tau)$をみたすもののこと

■部分単体の増大族$(K_r)_{r\in\mathbb{R}}, K_r := \{\sigma \in K \mid f(\sigma) \leq r\}$もフィルトレーションと呼ぶ




包含写像$K_r \hookrightarrow K_s\ (r \leq s)$から誘導されたホモロジーの族

$$PH_n(f)\colon \cdots \to H_n(K_r) \to H_n(K_s) \to \cdots$$

を$f$の（劣位集合フィルトレーションの）**パーシステントホモロジー**

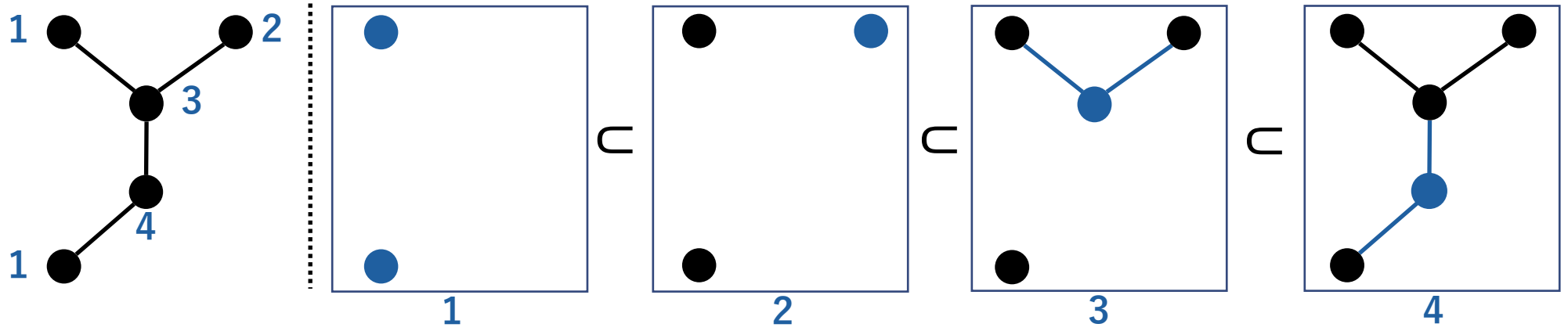
$$\begin{array}{lccccccccc} H_0\colon & 0 & \to & (\mathbb{F}_2)^3 & \to & (\mathbb{F}_2)^2 & \to & \mathbb{F}_2 & \to & \mathbb{F}_2 \\ H_1\colon & 0 & \to & 0 & \to & 0 & \to & \mathbb{F}_2 & \to & 0 \end{array}$$

# データからフィルトレーションを作る

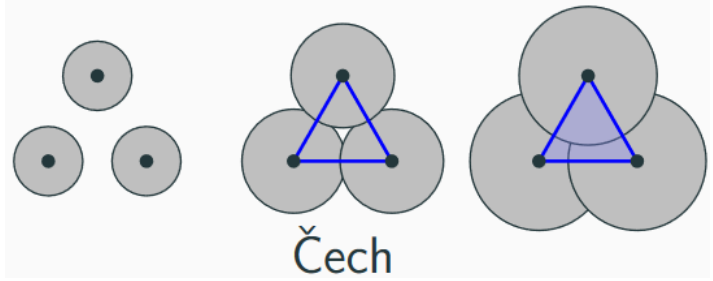
データから単体的複体のフィルトレーションを作ることができる

■単体的複体の頂点に関数$f$を与えると$f(\sigma) = \max_{v \in \sigma} f(v)$でフィルトレーションが定まる（劣位集合フィルトレーション）

■特に（無向単純）グラフの頂点上の関数からフィルトレーションが定まる




■Čech・VR・アルファ複体は$r$を動かすことでフィルトレーションを与える：$C(P) = \left(C(P;r)\right)_{r \in \mathbb{R}}$など

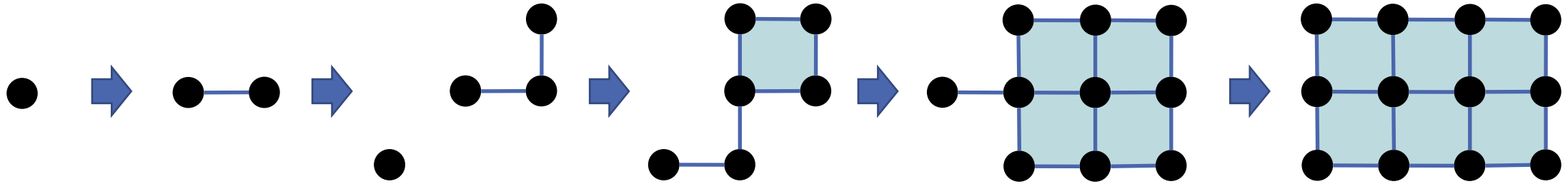


# 画像データからフィルトレーション

モノクロ2次元画像はグリッドに0~255の値が乗っているもの

| | | | |
|---|---|---|---|
| 255 | 200 | 65 | 35 |
| 200 | 155 | 0 | 25 |
| 255 | 35 | 65 | 190 |

**立方体的複体**とみなして，単体的複体と同様に劣位集合フィルトレーションが作れる



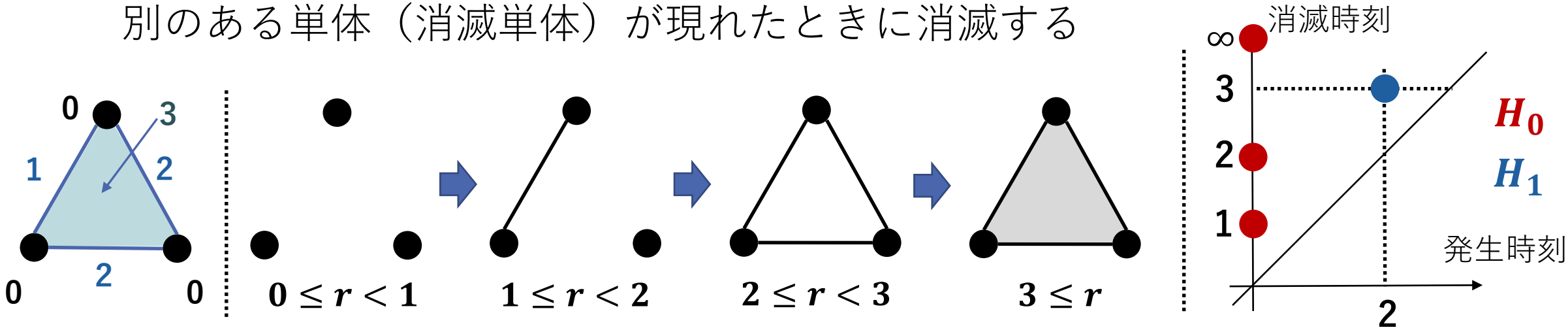
$0 \leq r < 25$　$25 \leq r < 35$　$35 \leq r < 65$　$65 \leq r < 155$　$200 \leq r < 255$　$255 < r$

モノクロ3次元画像も立方体的複体のフィルトレーションとみなせる

# パーシステンス図の計算方法

$f: K \to \mathbb{R}$ フィルトレーション

■各ホモロジー類は，ある単体（生成単体）が現れたときに生成し，
別のある単体（消滅単体）が現れたときに消滅する



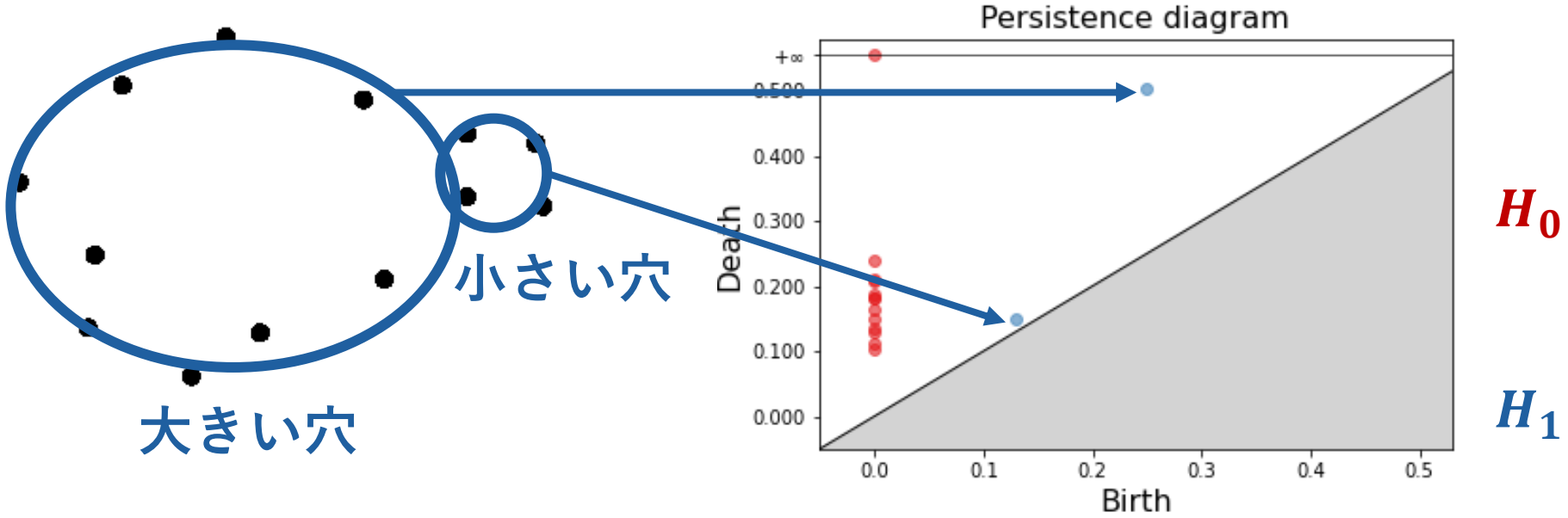

■上の例の1次のホモロジー類については　：生成単体　　：消滅単体

■パーシステンス図の点は生成単体$\sigma_b$と消滅単体$\sigma_d$のフィルトレーション値$f(\sigma_b)$と$f(\sigma_d)$を並べたもの（消滅単体がないときは∞）

■生成単体と消滅単体は境界準同形の行列表示の掃き出しで計算可能

# PDは何がうれしいのか

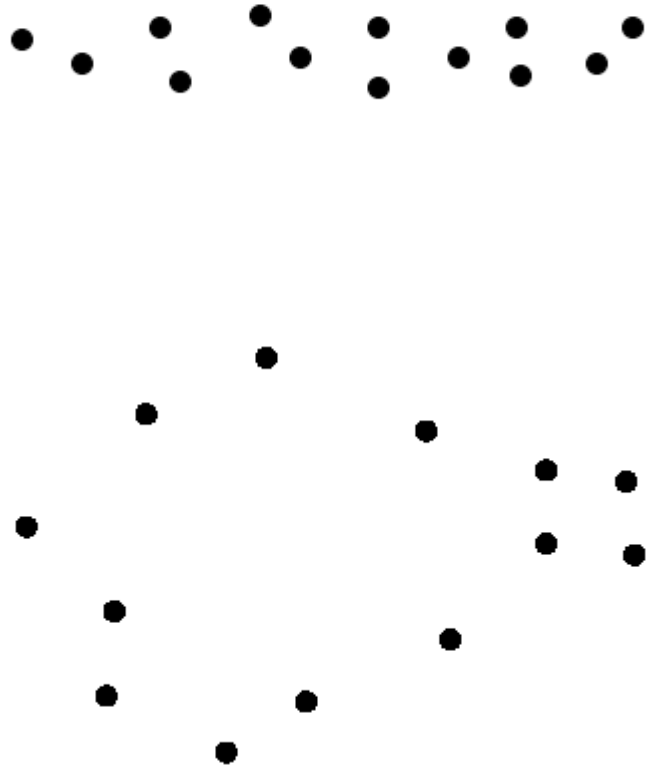
- ■**対角線から遠い点**（長く継続するホモロジー類）が**本質的な形**，**対角線の近くの点**（すぐに消滅するホモロジー類）は**ノイズ的なトポロジー**と区別できる
  - ■より一般にトポロジー的特徴のスケールもはかることができる
- ■パーシステンス図の点からホモロジー類を代表する「良い」サイクルを計算する手法がいろいろと提案されている（**逆解析**）

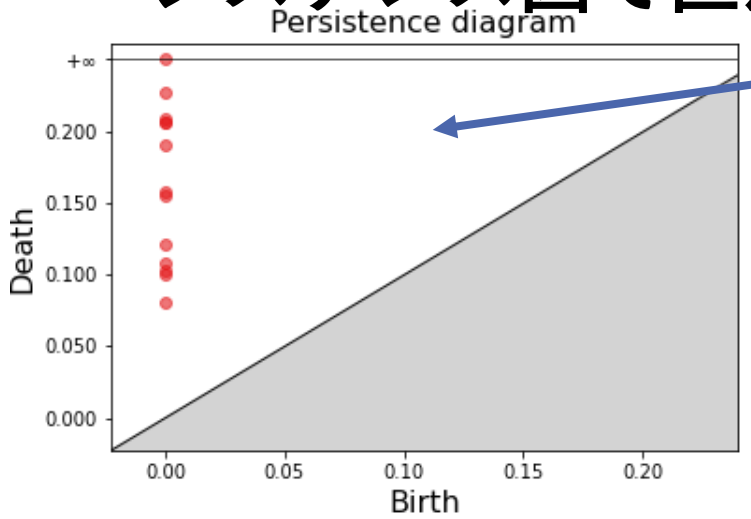


# PDは何がうれしいのか

もっと直接的には．．．

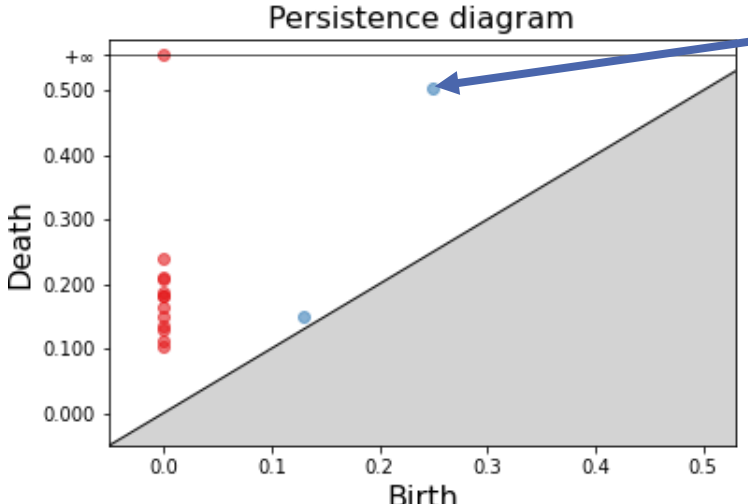
「**トポロジー**」**が異なるデータをパーシステンス図で区別**できる






**穴がない**

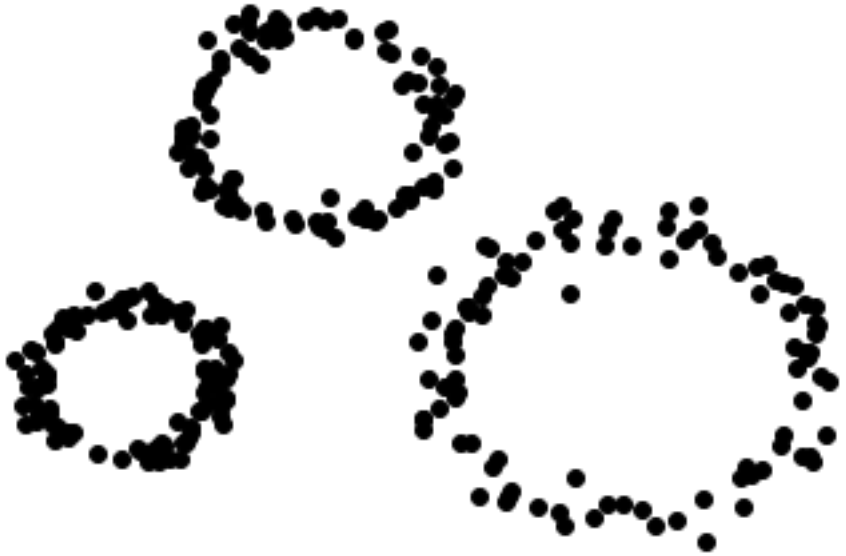
$H_0$



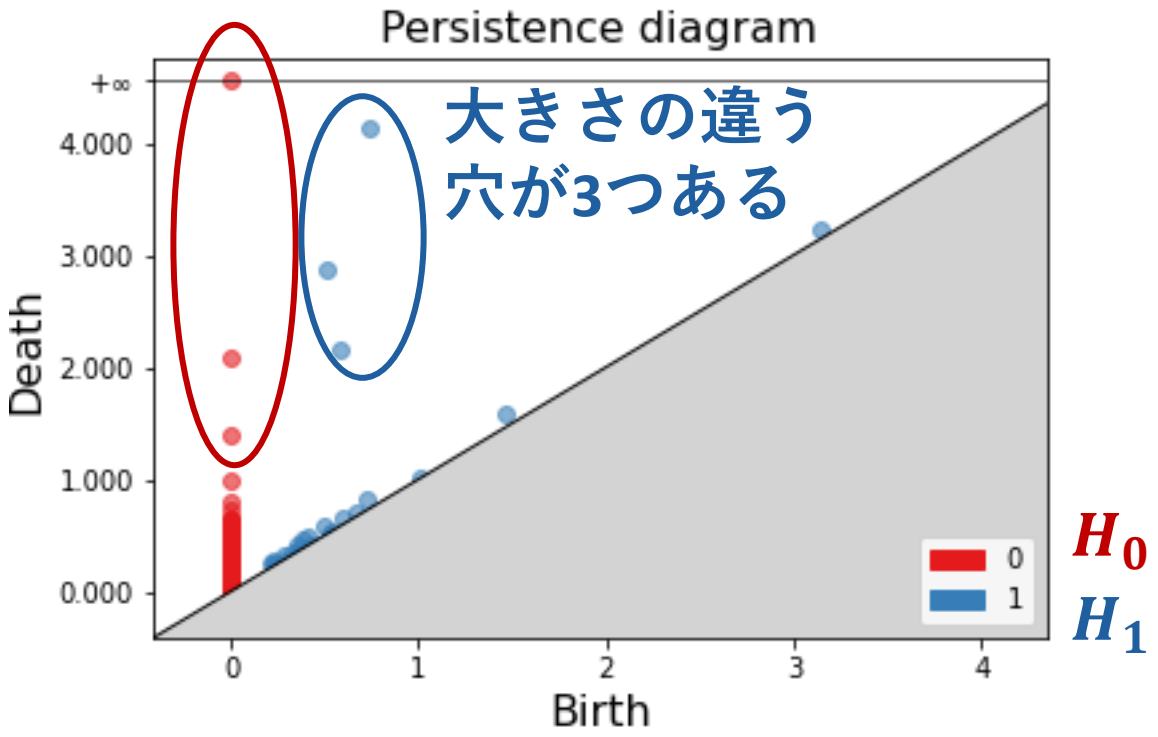

**穴がある**

$H_1$

# PHの簡単な使用例



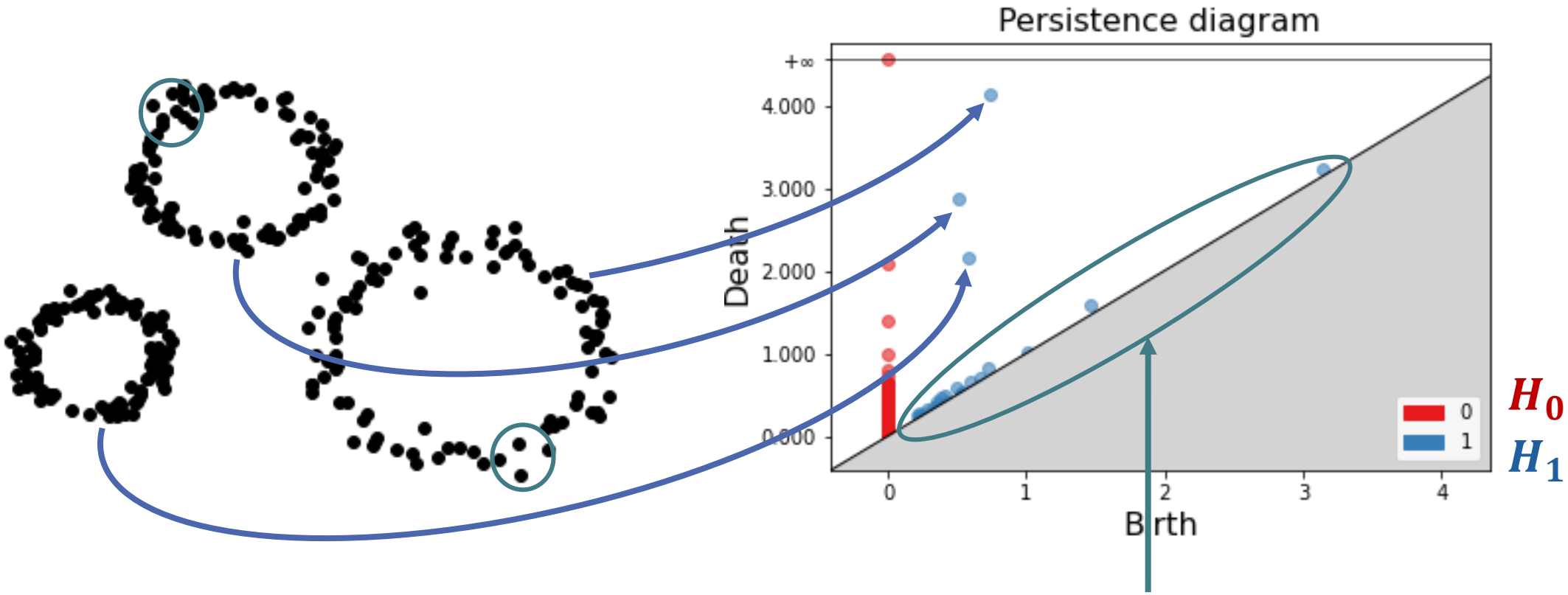
連結成分が3つある




$H_0$

$H_1$

# PHの簡単な使用例


Persistence diagram
+∞
4.000
3.000
2.000
1.000
0.000
Death
0
1
2
3
4
Birth
0
1
$H_0$
$H_1$
これらは生存時間が短い
のでノイズ由来と思える

# ボトルネック距離と安定性定理

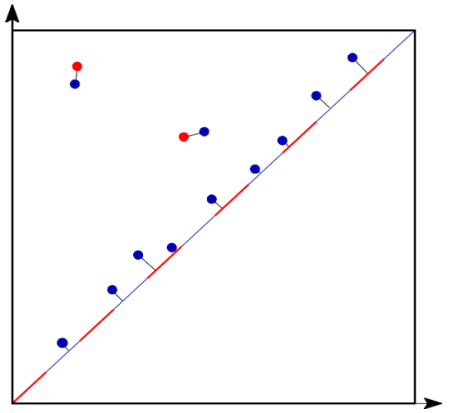
PD間の主要な距離のひとつ：**ボトルネック距離**

■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

$$d_B(D, D') \coloneqq \inf_{\gamma} \sup_{q \in D \cup \Delta} ||q - \gamma(q)||_\infty$$

$\gamma: D_1 \cup \Delta \to D_2 \cup \Delta$ ：全単射



**安定性定理**：ボトルネック距離を入力データの距離で上から評価

$f, g: X \to \mathbb{R}$としたとき，$d_B\big(D_n(f), D_n(g)\big) \leq \|f - g\|_\infty$

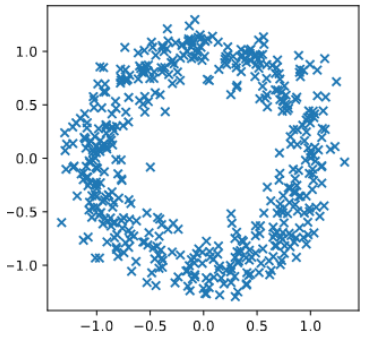
特に，有限集合$P, Q \subset \mathbb{R}^d$に対し，$d_B\big(D_n(C(P)), D_n(C(Q))\big) \leq d_H(P, Q)$

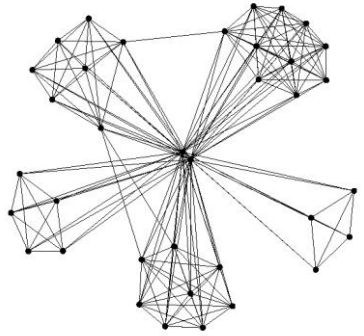
■現代的な理解ではPH間の代数的な距離を経由して証明する

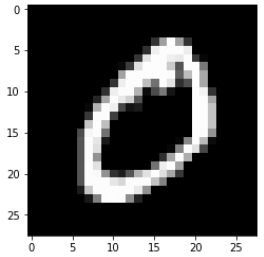
cf. [Cohen-Steiner–Edelsbrunner–Harer, 2007, Chazal–Cohen-Steiner–Glisse–Guibas–Oudot, 2009]

# TDAの典型的な使い方

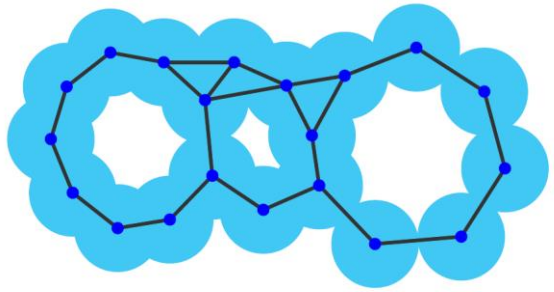
データ





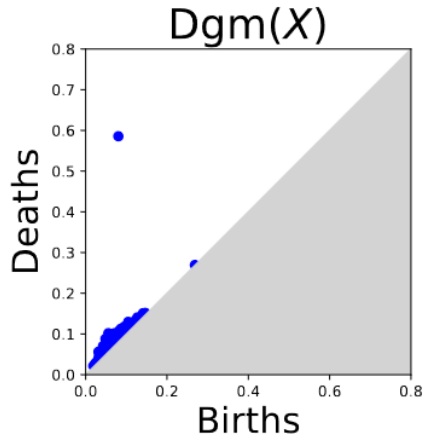


フィルトレーション

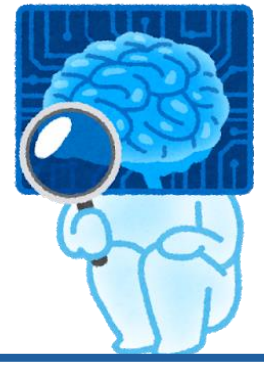


**Software**
Ripser, GUDHI,
HomCloud, …

パーシステンス図




専門家



機械学習

次回もう少し説明

# 位相的データ解析の応用例 パート1

点群データの解析，特に物質科学への応用

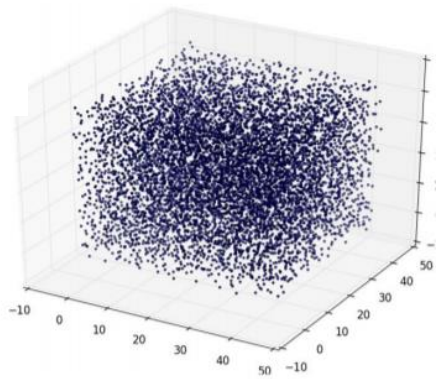
# シリカガラスの解析

**Nakamura et al. Persistent Homology and Many-Body Atomic Structure for Medium-Range Order in the Glass, 2015**

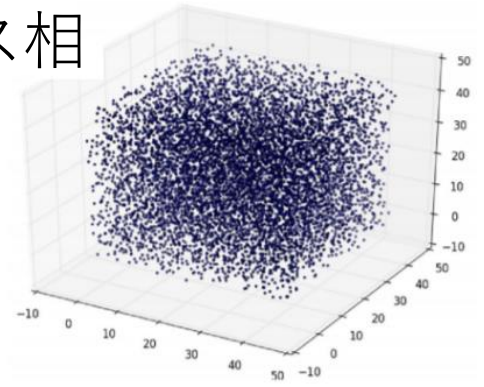
**$SiO_2$の液相とガラス相の違いをTDAを用いて明らかにした**

- ■$SiO_2$を液相から急速に冷却するとガラス相になる
- ■シミュレーションで原子配置を用意する

液相



ガラス相



**■原子配置を見るだけでは差が微妙で区別が困難**

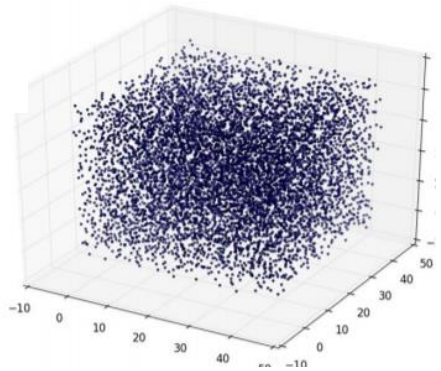
Figures from K. Fukumizu, Persistence Weighted Gaussian Kernel for Topological Data Analysis

# シリカガラスの解析

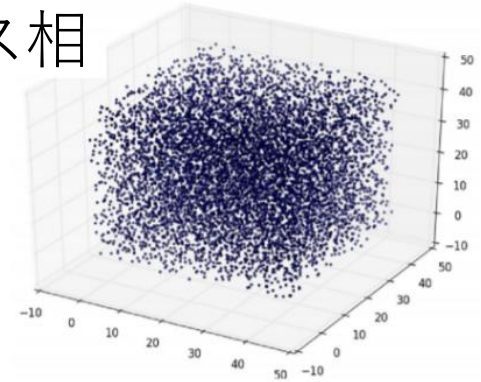
**Nakamura et al. Persistent Homology and Many-Body Atomic Structure for Medium-Range Order in the Glass, 2015**

アイデア：**点群をPDに変換**して，それらを調べる

液相

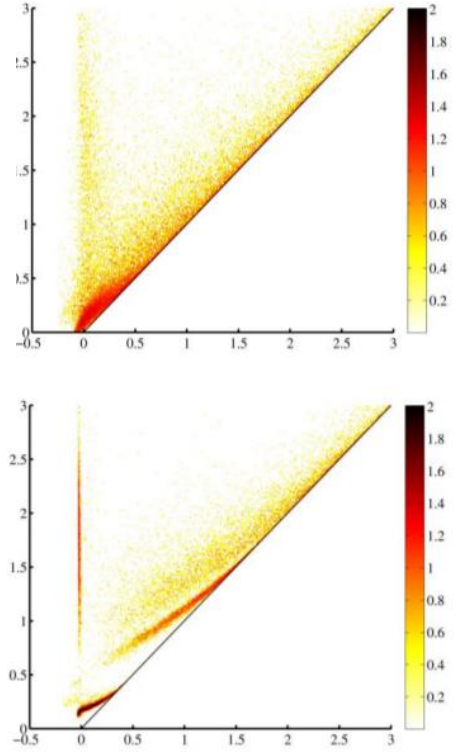


ガラス相



PD

$H_1$



※点が多いのでPDを密度で表示

※SiとOで球の半径を変えて計算

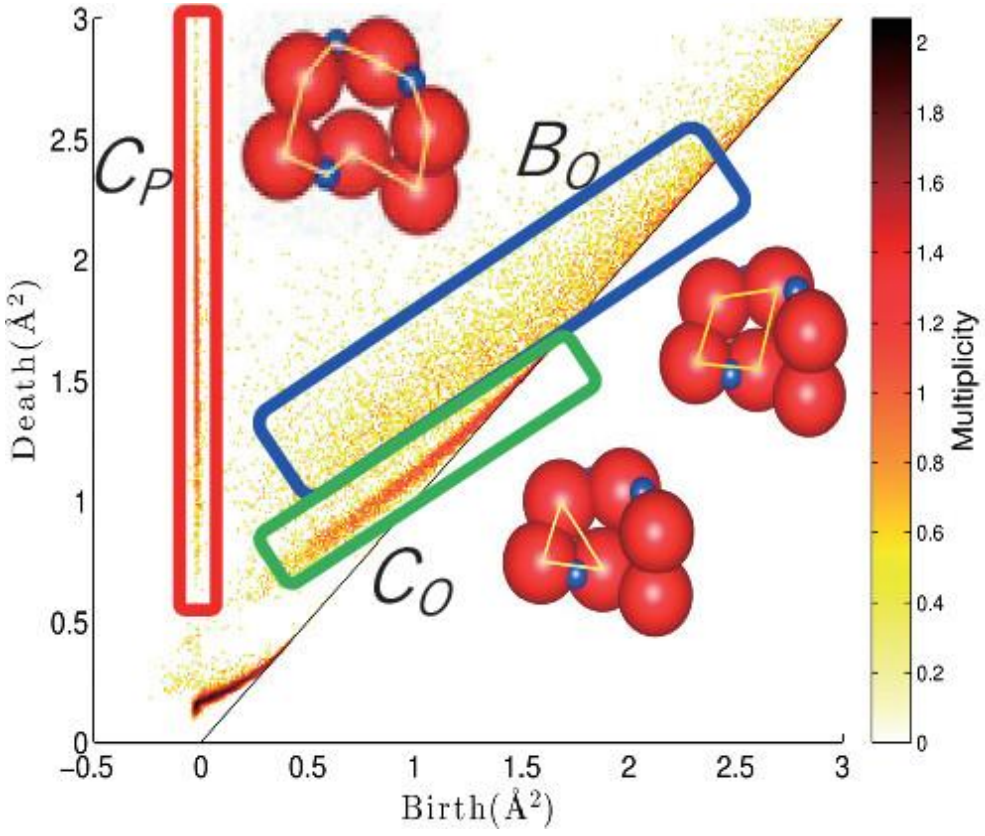
**パーシステンス図に変換すると差が明らかに見える**

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

# シリカガラスの解析：PDの意味

$SiO_2$のガラス相のPDに特徴的な曲線の意味を調べる

**PDの点が**原子配置点群の**どの形に対応するかを出力可能（逆解析）**



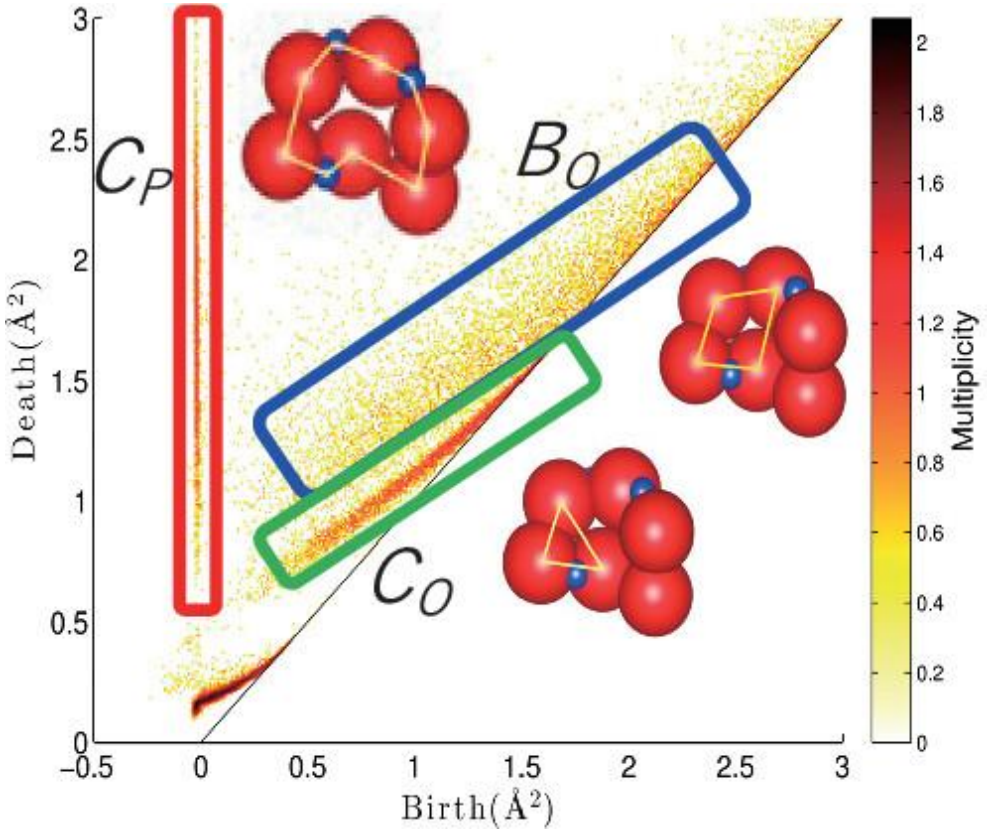

- $C_P$: ⋯-O-Si-O-Si-O-⋯という共有結合によるリング
- $C_O$: Si原子をまたぐ3つのO原子が作る三角形
- $B_O$: いろいろな構造が混在
  たとえば，Si原子をまたいだ4個以上のO原子のなす多角形

Figure from Obayashi et al., Persistent homology analysis for materials research and persistent homology software: Homcloud

# シリカガラスの解析：PDの意味

$SiO_2$のガラス相のPDに特徴的な曲線の意味を調べる

**PDの点が**原子配置点群の**どの形に対応するかを出力可能（逆解析）**




- 結晶は長距離秩序（繰り返し構造）を持つがガラスは持たない
- 結晶・ガラス・液体はすべて短距離秩序（隣接原子に関する空間パターン）を持つ
- ガラスは液体にはない**中距離秩序**を持つと考えられているが，$C_O, B_O, C_P$がそれに対応すると信じられている

**PDがデータの新しい記述子となる可能性**

Figure from Obayashi et al., Persistent homology analysis for materials research and persistent homology software: Homcloud

# 金属ガラスの解析

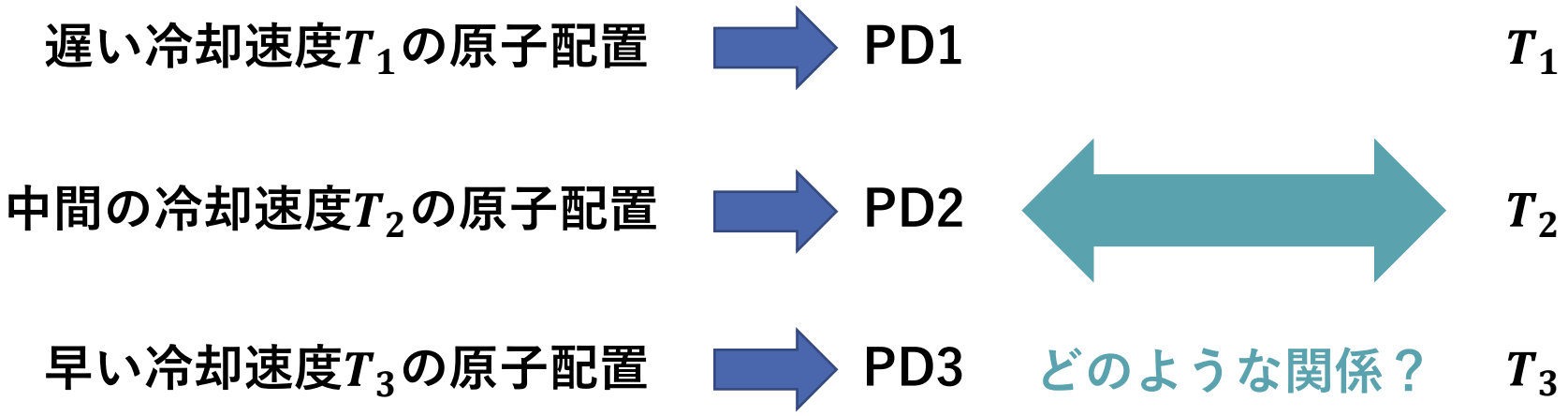
**Hirata et al., Structural changes during glass formation extracted by computational homology with machine learning, 2020**

Pd 80%とSi 20%からなる**金属ガラス**の解析

- ■高速冷却のシミュレーションで原子配置を用意
- ■複数の冷却速度でデータを生成
- ■原子配置のPDと冷却速度の関係を調べた

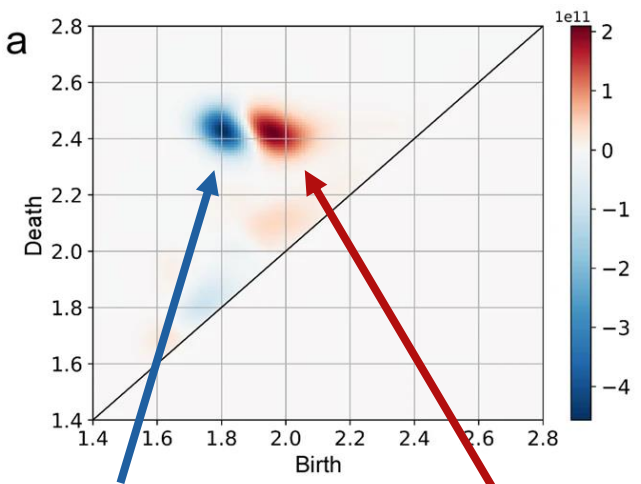
**次回もう少し説明**

（ベクトル化+線形回帰）

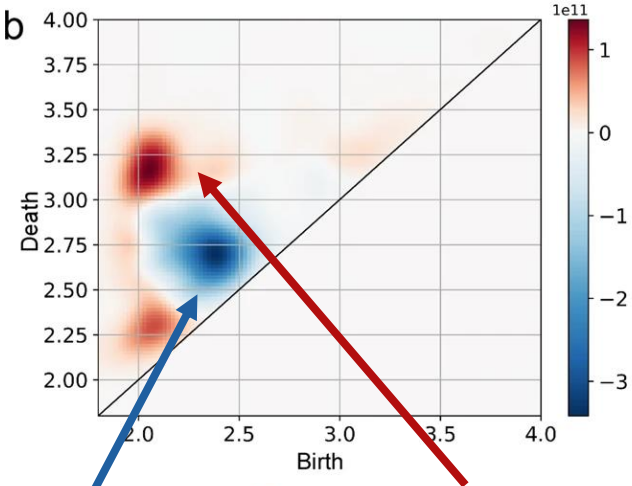


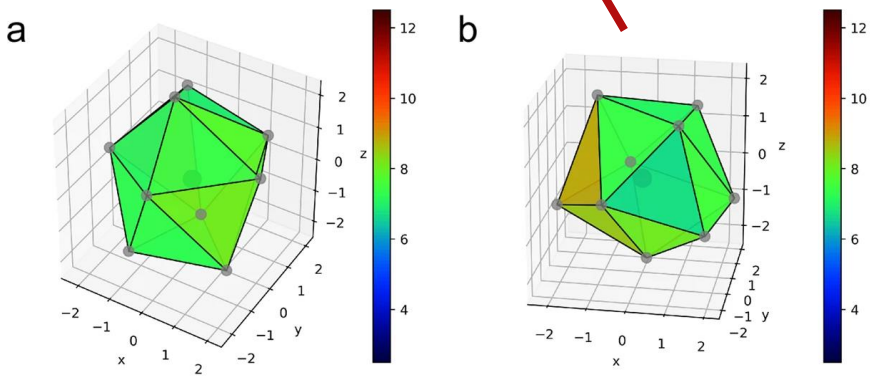
# 金属ガラスの解析

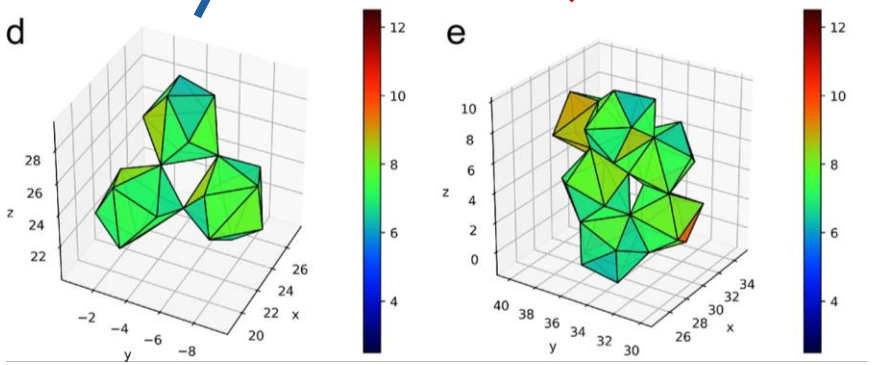
Pd原子配置の$H_2$に関するPD




Si原子配置の$H_1$に関するPD




冷却温度が早いほど，赤い領域の点が多く，青い領域の点が少ない

■具体的にどの原子配置の構造が冷却速度に関係するかもわかる

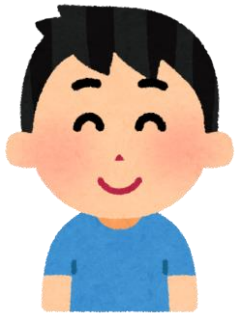
# せん妄検知への応用

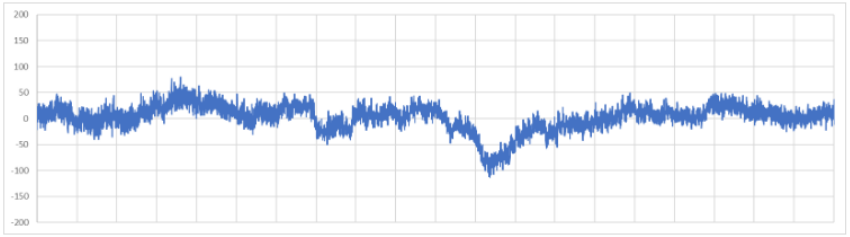
**Kajitani et al. Application of Topological Data Analysis to Delirium Detection, 2020**

**せん妄**：身体に負担がかかったときに生ずる脳機能の乱れ

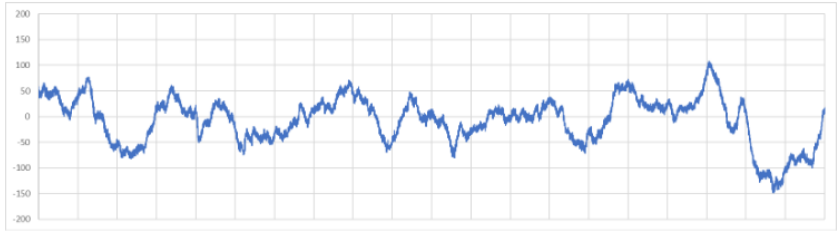
入院患者に生じることが多い

↝ 脳波からせん妄を判定できるとうれしい





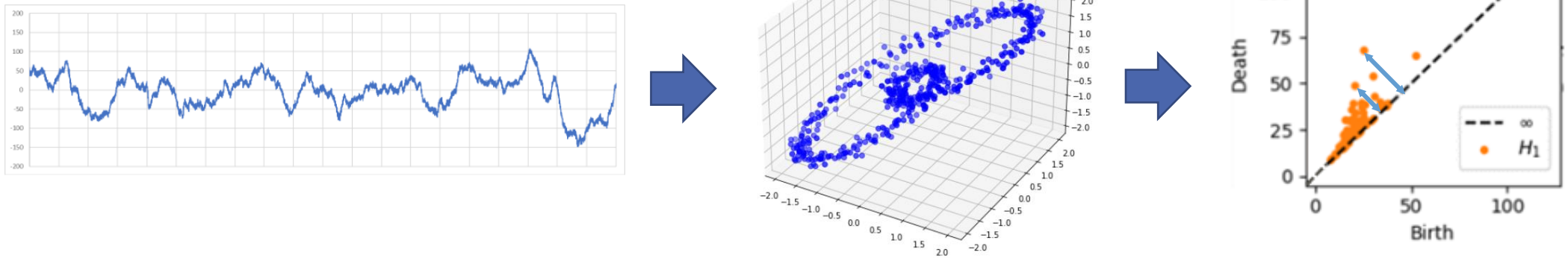
正常時の脳波



せん妄時の脳波

従来のフーリエ解析的手法と位相的データ解析の手法を組み合わせて使うことで，高精度に脳波からせん妄を検知

# せん妄検知への応用：手法




1. 脳波を時間遅延埋め込みで**3次元点群に変換**

$$[x_0, x_1, x_2, x_3, x_4 \ldots] \mapsto \left\{ \begin{bmatrix} x_0 \\ x_1 \\ x_2 \end{bmatrix}, \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}, \begin{bmatrix} x_2 \\ x_3 \\ x_3 \end{bmatrix}, \ldots \right\}$$ この点群の差で脳波の差を検知可

2. 位相的データ解析を用いて**PDに変換**（$H_1$を使う）
3. PDの点たちの**対角線からの距離の和でスコアを計算**

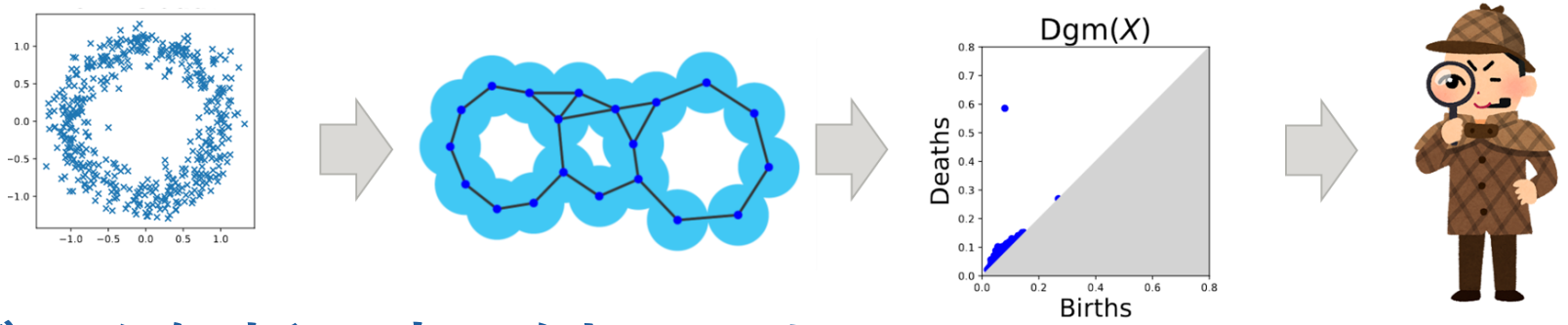
| | TDA-EEG | Shinozaki *et al.* [10] | Numan *et al.* [6] | Adaboost |
|---|---|---|---|---|
| AUC | **0.80** | 0.72 | 0.66 | 0.69 |
| Specificity (Sensitivity = 0.75) | **0.71** | 0.52 | 0.42 | 0.50 |

従来手法に比べて**見逃し率が約半減**

# まとめ

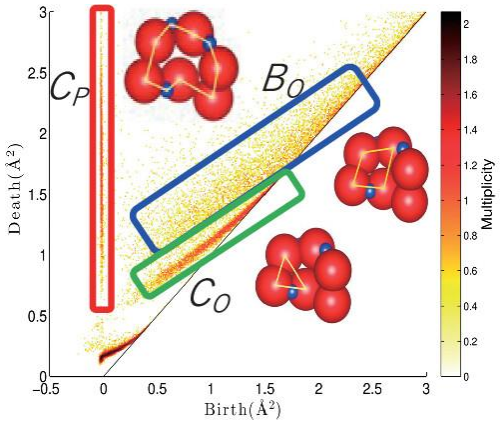
## 1. パーシステントホモロジー

- データの「トポロジー」をマルチスケールに抽出する手法
- フィルトレーションを定めて**パーシステンス図**を計算
- データの摂動に対して頑健であることを保証する**安定性定理**が成立




## 2. 位相的データ解析の応用例 パート1

- **物質科学**に対して有用に応用された
- パーシステンス図の点に対して**逆解析**ができることが強み




■次回は**機械学習との組み合わせ**について説明したい

# 参考文献


- H. Edelsbrunner and J.L. Harer, Computational Topology: An Introduction, American Mathematical Society, 2010.
  （翻訳版：荒井迅・竹内博志訳，計算トポロジー入門，2023）
- 平岡裕章，タンパク質構造とトポロジー ―パーシステントホモロジー群入門―，共立出版，2013.
- S.Y. Oudot, Persistence Theory: From Quiver Representations to Data Analysis, American Mathematical Society, 2015.
- 池祐一・E.G. エスカラ・大林一平・鍛冶静雄，位相的データ解析から構造発見へ：パーシステントホモロジーを中心に，サイエンス社，2023.