Kernel Methods for Topological Data Analysis

Kenji Fukumizu

The Institute of Statistical Mathematics (Tokyo, Japan)

Joint work with Genki Kusano and Yasuaki Hiraoka (Tohoku Univ.), supported by JST CREST

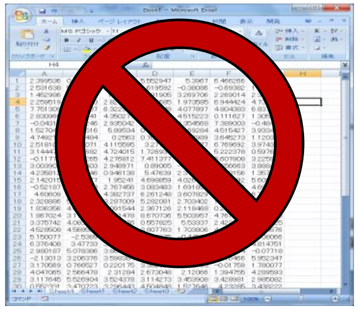
STM2016 at ISM. July 22, 2016

Topological Data Analysis

- TDA: a new method for extracting topological or geometrical information of data.

Key technology = Persistence homology
(Edelsbrunner et al 2002; Carlsson 2005)

Background
- Complex data:
  Data with complex structure must be analyzed.
- Progress of computational topology:
  Computing topological invariants becomes easy.



TDA: Various applications

Data of highly complex geometric structure

Often difficult to define good feature vectors / descriptors

Brain Science

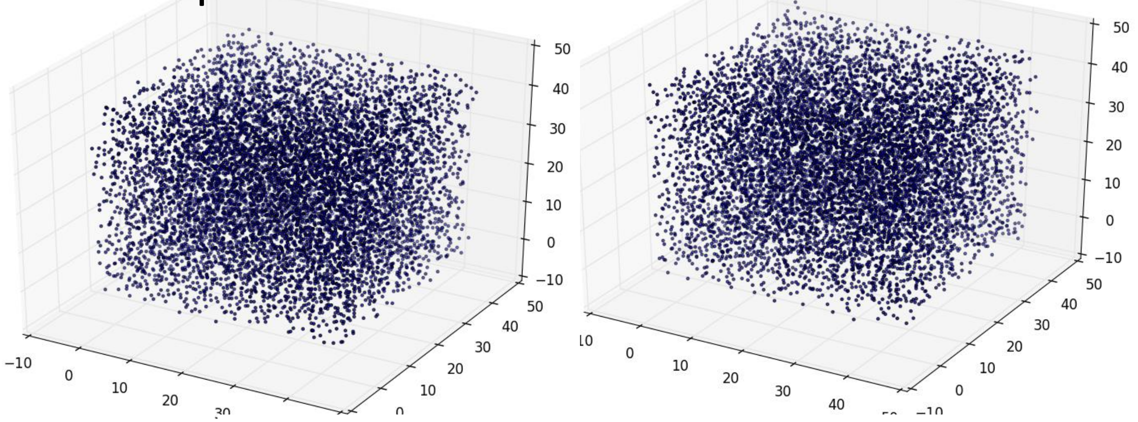
Brain artery trees
e.g. age effect
(Bendich et al 2014)

Biochemistry

Structure change
of proteins
eg. open / closed
(Kovacev-Nikolic et al 2015)

Material Science

Liquid Glass



Non-crystal materials
(Nakamura, Hiraoka, Hirata, Escolar, Nishiura. *Nanotechnology* 26 (2015))

Computer Vision

Shape signature,
natural image statistics
(Freedman & Chen 2009)

etc…

Persistence homology provides a compact representation for such data.

Outline

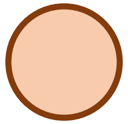
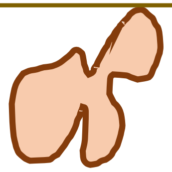
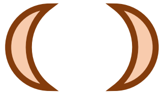
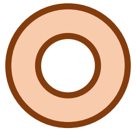
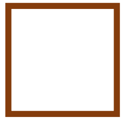
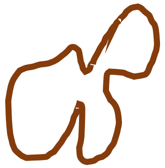
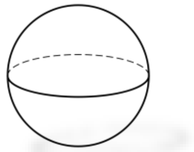
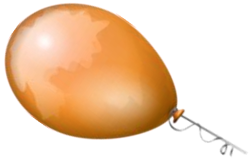
- A brief introduction to persistence homology
- Statistical approach with kernels to topological data analysis
- Applications
  - Material science
  - Protein classification
- Summary

Topology

$\cong$

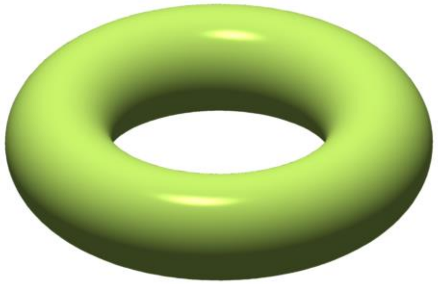
Topology: two sets are equivalent if one is deformed to the other without tearing or attaching.

Topological invariants: any equivalent sets take the same value.

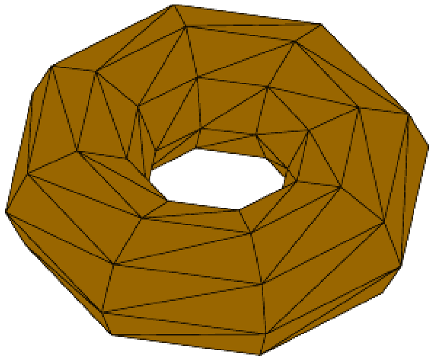
	Connected components	Ring	Cavity
 $\cong$  $\cong$ 	1	0	0
 $\cong$ 	2	0	0
 $\cong$  $\cong$ 	1	1	0
 $\cong$ 	1	0	1

Algebraic Topology

- Algebraic treatment of topological spaces



$\cong$



Simplicial complex
(union of simplexes)

Algebraic operations

Compute various topological invariances
e.g. Euler number

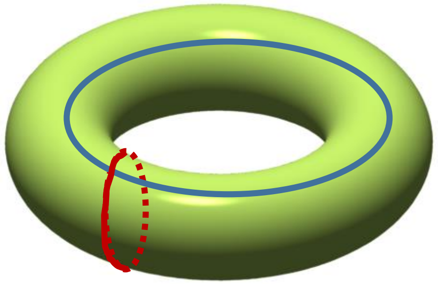
Classify topological spaces with topological invariances.

- Homology group: independent “holes”

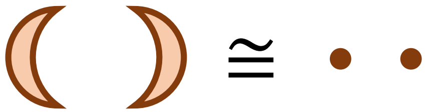
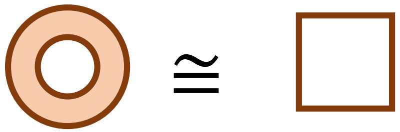
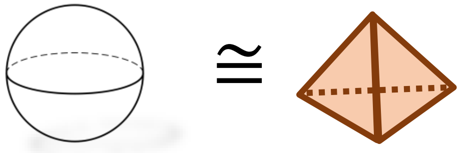
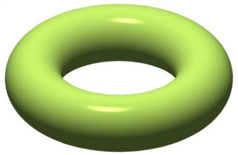
$H_k(X)$: k-th homology group of topological space X ($k = 0,1,2,\ldots$)

k-dimensional holes

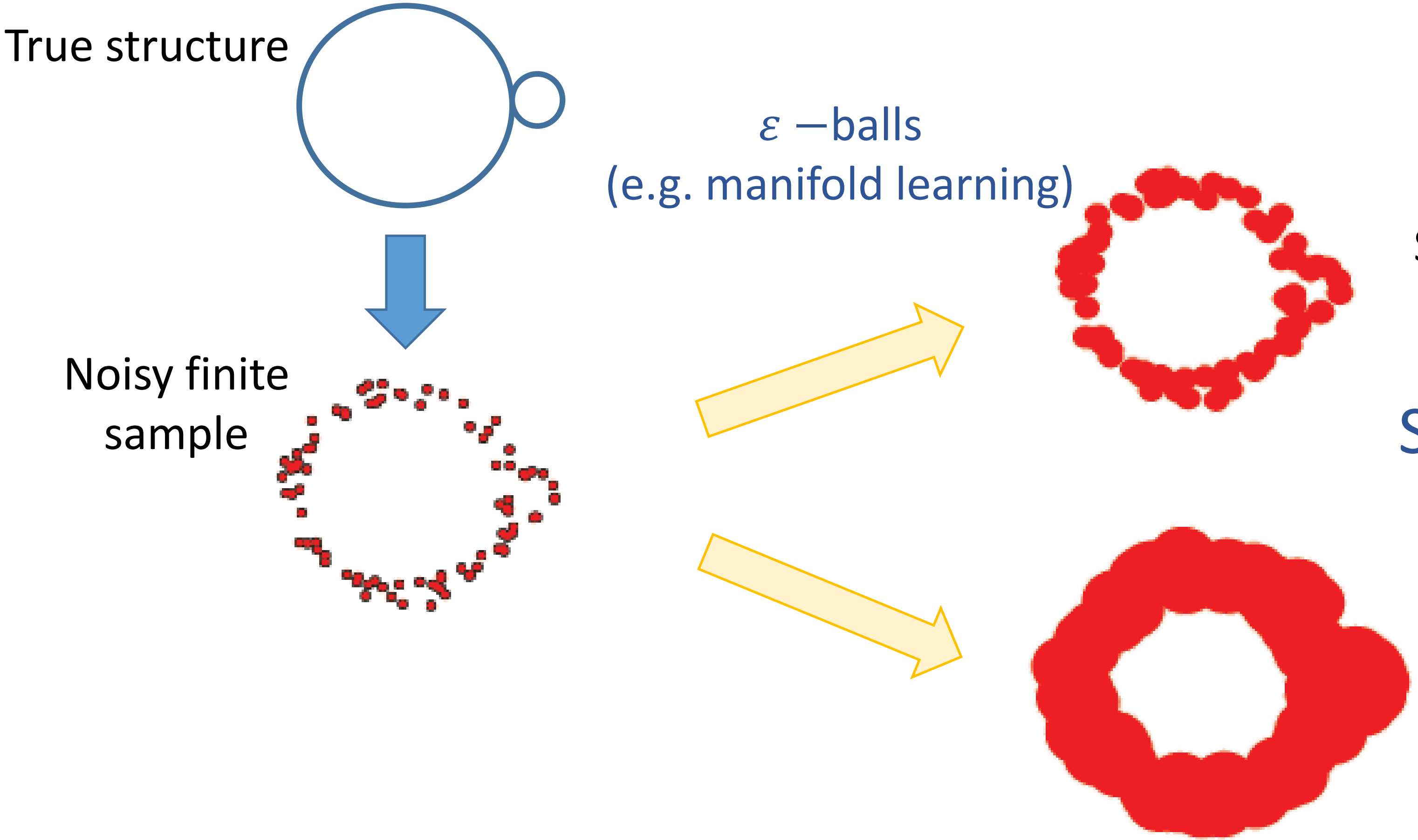
$H_0(X)$: connected components
$H_1(X)$: rings
$H_2(X)$: cavities
…



The generators of 1st homology group

	$H_0(X)$	$H_1(X)$	$H_2(X)$
	$\mathbb{Z}$	0	0
	$\mathbb{Z} \oplus \mathbb{Z}$	0	0
	$\mathbb{Z}$	$\mathbb{Z}$	0
	$\mathbb{Z}$	0	$\mathbb{Z}$
	$\mathbb{Z}$	$\mathbb{Z} \oplus \mathbb{Z}$	$\mathbb{Z}$

Topology of statistical data?



Small ε → disconnected object

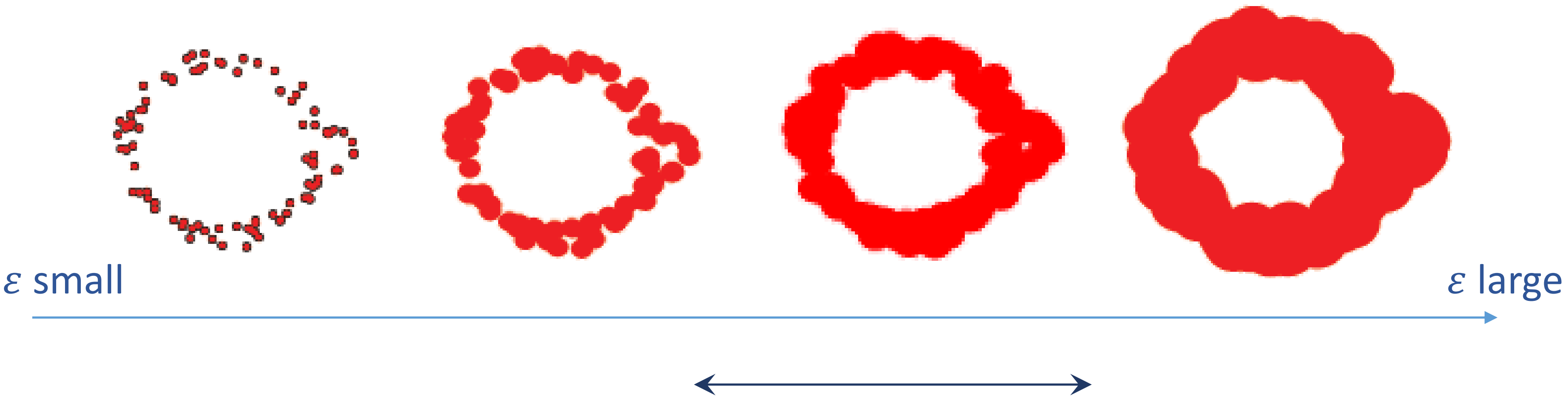
Stable extraction of topology is NOT easy!

Large ε → small ring is not visible

Persistence Homology

- All ε considered

$$X = \{x_i\}_{i=1}^{m} \subset \mathbf{R}^d , \quad X_\varepsilon := \cup_{i=1}^{m} B_\varepsilon(x_i)$$



Two rings （generators of 1 dim homology）
persist in a long interval.

- Persistence homology (formal definition)

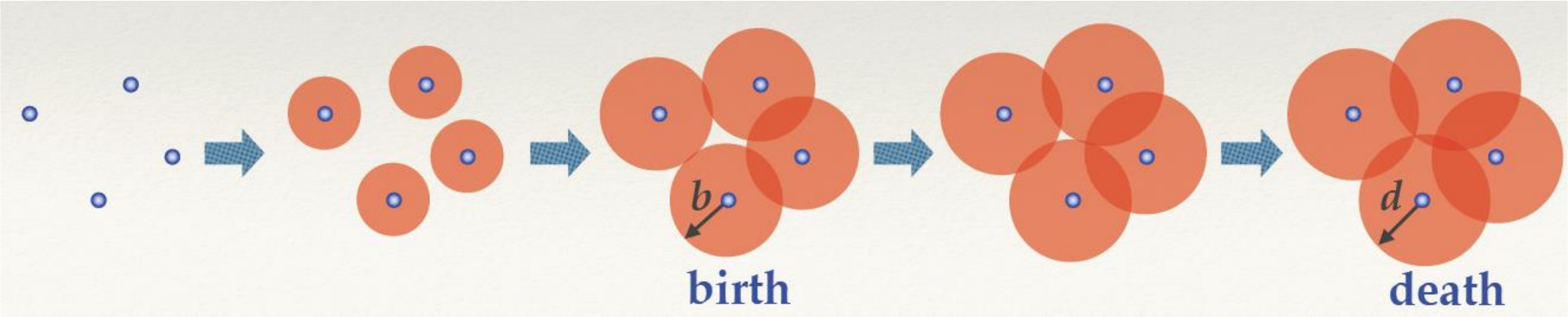
Filtration of topological spaces $\mathfrak{X}: \; X_1 \subset X_2 \subset \cdots \subset X_L$

$$PH_k(\mathfrak{X}): H_k(X_1) \to H_k(X_2) \to \cdots \to H_k(X_L) \;\cong \bigoplus_{i=1}^{m_k} I[b_i, d_i]$$ Irreducible decomposition

$$I[b, d\,] \cong 0 \to \cdots \to 0 \to \overset{\text{at } X_b}{K} \to \cdots \to \overset{\text{at } X_d}{K} \to 0 \to \cdots \to 0$$ K: field

The lifetime (birth, death) of each generator is rigorously defined, and can be computed numerically.

Birth and death of a generator of $PH_1(X)$

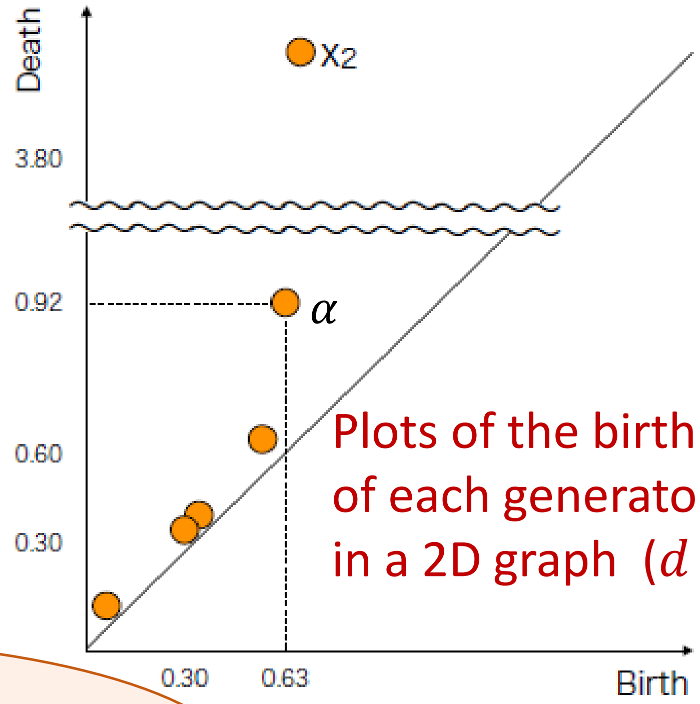


- Two popular (equivalent) expressions of PH

Barcode

Bar from the birth to death of each generator

Persistence diagram (PD)



Plots of the birth (b) and death (d) of each generator of PH in a 2D graph ($d \geq b$).

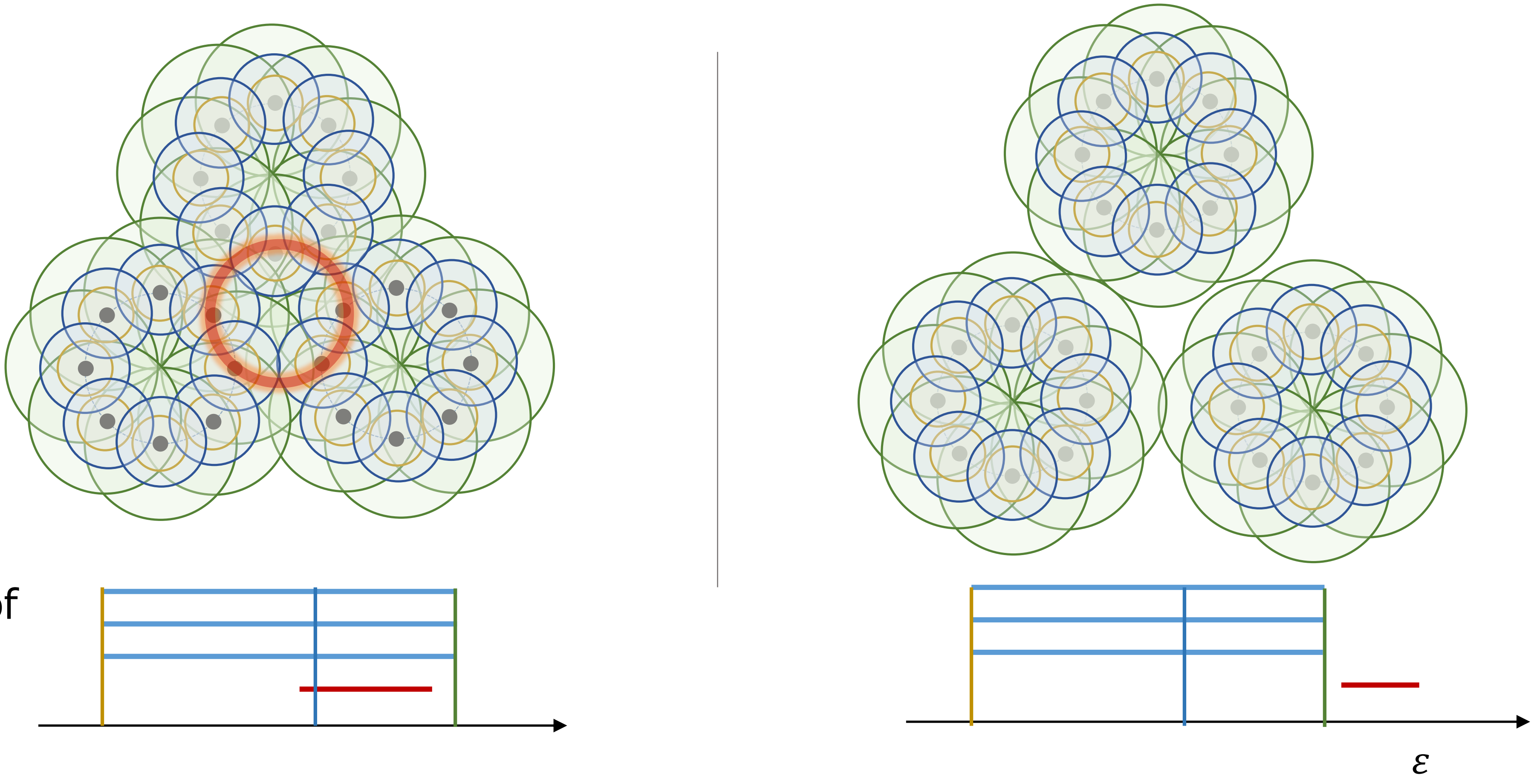
Handy descriptors or features of complex geometric objects

Barcodes and PD are considered for each dimension.

Beyond topology

- PH contains geometrical information more than topology

Barcodes of
1-dim PH

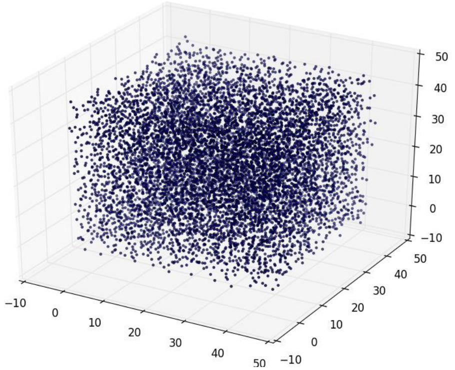


Statistical approach with kernels to topological data analysis

Statistical approach to TDA

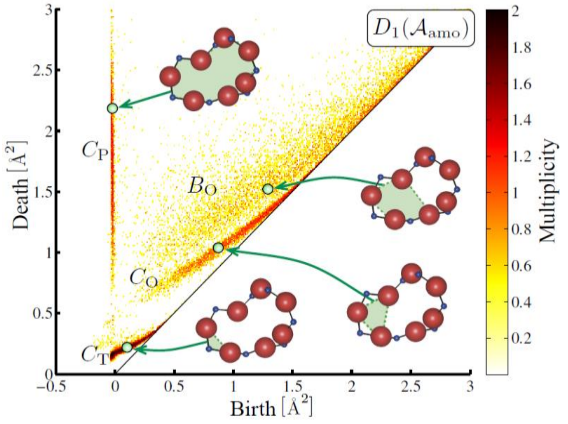
- <u>Conventional</u> TDA

Data | Computation of PH | Visualization（PD） | Analysis by experts



e.g. Molecular dynamics simulation

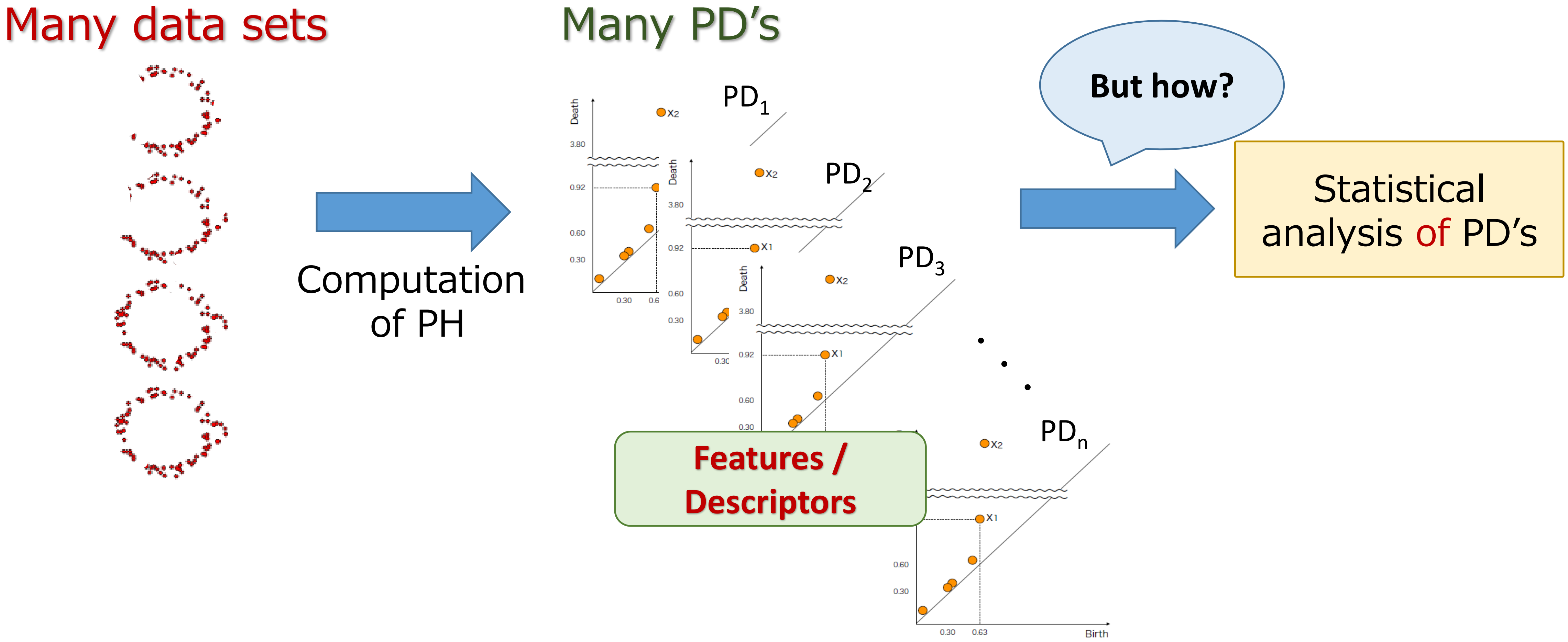
Software
CGAL / PHAT



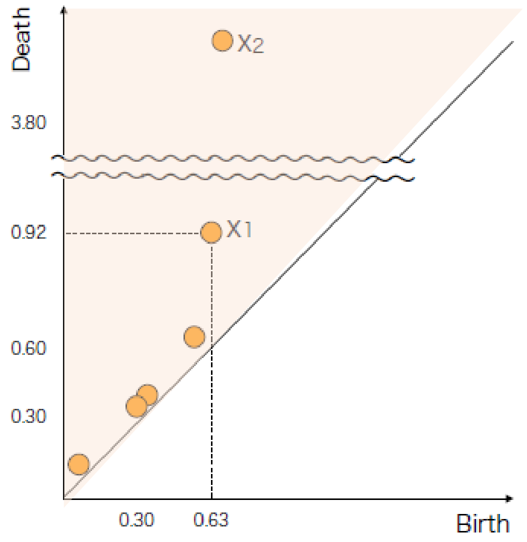
CGAL: The Computational Geometry Algorithms Library http://www.cgal.org/
PHAT: Persistent Homology Algorithm Toolbox https://bitbucket.org/phat-code/phat

- Statistical approach to TDA

(Kusano, Fukumizu, Hiraoka ICML 2016; Reininghaus et al CVPR 2015; Kwitt et al NIPS2015; Fasy et al 2014)



Kernel representation of PD



- Vectorization of PD by positive definite kernel
  - PD = Discrete measure $\quad \mu_D := \sum_{z \in PD} \delta_z$

- Kernel embedding of PD's into RKHS

$$\mathcal{E}_k: \quad \mu_D \mapsto \int k(\cdot, x) d\mu_D(x) = \sum_i k(\cdot, x_i) \quad \in H_k, \quad \text{Vectorization}$$

  - For some kernels (e.g., Gaussian, Laplace), $\mathcal{E}_k$ is injective.

k: positive definite kernel
H_k: corresponding RKHS

- By vectorization,
  - a number of methods for data analysis can be applied,
    SVM, regression, PCA, CCA, etc.
  - tractable computation is possible with kernel trick.

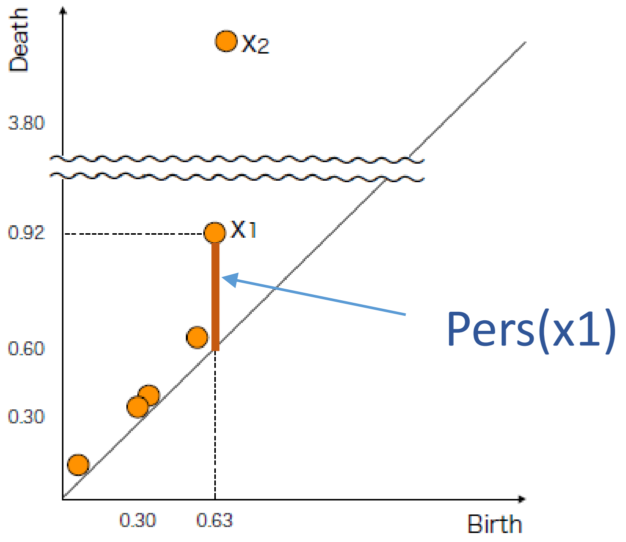
Persistence Weighted Gaussian (PWG) Kernel

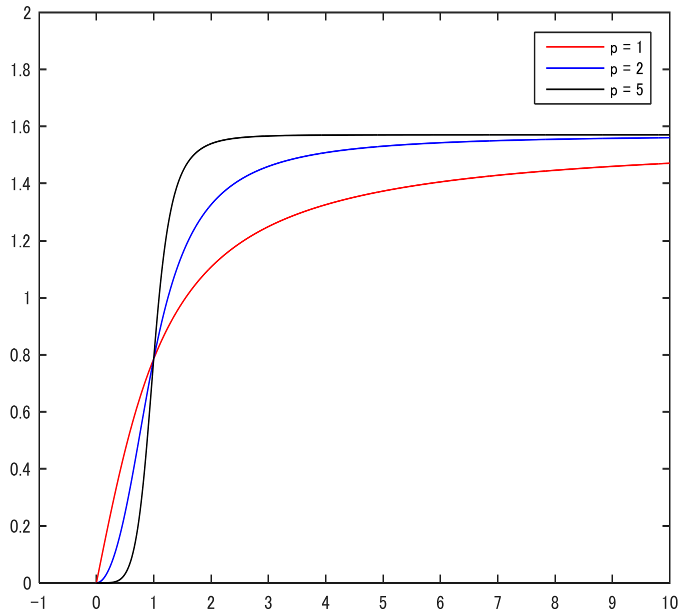
Generators close to the diagonal may be noise, and should be discounted.

$$k_{PWG}(x,y) = w(x)w(y)\exp\left(-\frac{\|y-x\|^2}{2\sigma^2}\right)$$

$$w(x) = w_{C,p}(x) := \arctan(C\,\mathrm{Pers}(x)^p) \qquad (C, p > 0)$$

$$\mathrm{Pers}(x) := d - b \text{ for } x \in \{(b,d) \in \mathbf{R}^2 | d \geq b\}$$





- Stability with PWG kernel embedding
  - PWGK defines a distance on the persistence diagrams,

$$d_k(D_1, D_2) \coloneqq \|\mathcal{E}_k(D_1) - \mathcal{E}_k(D_2)\|_{H_k}, \qquad D_1, D_2\text{: persistence diagrams}$$

<u>Stability Theorem</u> (Kusano, Hiraoka, Fukumizu 2015)

M: compact subset in $\mathbf{R}^d$. $S \subset M$, $T \subset \mathbf{R}^d$: finite sets.

If $p > d + 1$, then with PWG kernel (p, C, σ),

$$d_k\big(D_q(S), D_q(T)\big) \leq L\, d_H(S, T).$$

A small change of a set causes only a small change in PD

Lipschitz continuity

L: constant depending only on M, p, d, C, σ

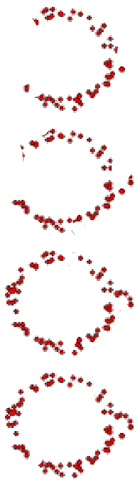
$D_q(S)$: q th persistence diagram of S

d_H: Haussdorff distance

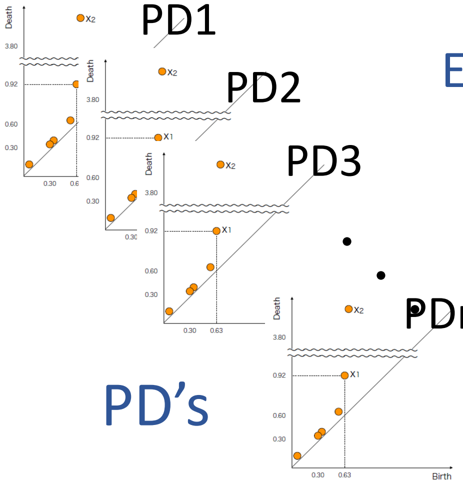
This stability is NOT known for Gaussian kernel.

2nd-level kernel

Data analysis method



Data sets



PD1

PD2

PD3

PDm

PD's

Embedding

$$\mathcal{E}_k(PD_1)$$
$$\mathcal{E}_k(PD_2)$$
$$\cdots$$
$$\mathcal{E}_k(PD_m)$$

Vectors in RKHS

Application of pos. def. Kernel on RKHS

2nd-level kernel (SVM for measures, Muandet, Fukumizu, Dinuzzo, Schölkopf 2012)

- RKHS-Gaussian kernel $K(\varphi_1, \varphi_2) = \exp\left(-\frac{\|\varphi_1 - \varphi_2\|_{H_k}^2}{2\tau^2}\right)$

  derives

$$K(D_i, D_j) = \exp\left(-\frac{\|\mathcal{E}_k(D_i) - \mathcal{E}_k(D_j)\|_{H_k}^2}{2\tau^2}\right)$$

D_i, D_j: Persistence diagrams

Computational issue

The number of generators in a PD may be large ($\geq 10^3, 10^4$)

For $PD_i = \sum_{a=1}^{N_i} \delta_{x_a^{(i)}} \cup \Delta$, $K(PD_i, PD_j) = \exp\left(-\frac{\|\mathcal{E}_k(PD_i)-\mathcal{E}_k(PD_j)\|_{H_k}^2}{2\tau^2}\right)$ requires computation

$$\|\mathcal{E}_k(PD_i) - \mathcal{E}_k(PD_j)\|_{H_k}^2$$
$$= \sum_{a=1}^{N_i} \sum_{b=1}^{N_i} k\left(x_a^{(i)}, x_b^{(i)}\right) + \sum_{a=1}^{N_j} \sum_{b=1}^{N_j} k\left(x_a^{(j)}, x_b^{(j)}\right) - 2\sum_{a=1}^{N_i} \sum_{b=1}^{N_j} k\left(x_a^{(i)}, x_b^{(j)}\right).$$

The number of $\exp\left(-\frac{\|x_a - x_b\|^2}{2\sigma^2}\right) = O(m^2 N^2)$ → computationally expensive for $N \approx 10^4$

$$N = \max\{N_i | i = 1, \dots, n\}$$

- Approximation by random features (Rahimi & Recht 2008)

By Bochner's theorem

Gaussian distribution $=: Q_\sigma$

$$\exp\left(-\frac{\|x_a-x_b\|^2}{2\sigma^2}\right) = C\int e^{\sqrt{-1}\omega^T(x_a-x_b)} \left(\frac{\sigma^2}{2\pi}\right) \boxed{e^{-\frac{\sigma^2\|\omega\|^2}{2}}\, d\omega}$$ (Fourier transform)

Approximation by sampling: $\omega_1, \ldots, \omega_L: i.i.d. \sim Q_\sigma$

$$\exp\left(-\frac{\|x_a-x_b\|^2}{2\sigma^2}\right) \approx C\frac{1}{L}\sum_{\ell=1}^{L} e^{\sqrt{-1}\omega_\ell^T x_a}\ \overline{e^{\sqrt{-1}\omega_\ell^T x_b}}$$

$$\sum_{a=1}^{N_i}\sum_{b=1}^{N_j} k\left(x_a^{(i)}, x_b^{(j)}\right) \approx \frac{C}{L}\sum_{a=1}^{N_i}\sum_{b=1}^{N_j}\sum_{\ell=1}^{L} w\left(x_a^{(i)}\right) w\left(x_b^{(j)}\right) e^{\sqrt{-1}\omega_\ell^T x_a^{(i)}}\ \overline{e^{\sqrt{-1}\omega_\ell^T x_b^{(j)}}}$$

$$= \frac{C}{L}\sum_{\ell=1}^{L} \underbrace{\boxed{\sum_{a=1}^{N_i} w\left(x_a^{(i)}\right) e^{\sqrt{-1}\omega_\ell^T x_a^{(i)}}}}_{L \text{ dim.}}\ \overline{\sum_{b=1}^{N_j} w\left(x_b^{(j)}\right) e^{\sqrt{-1}\omega_\ell^T x_b^{(j)}}}$$

Computational cost $O(LN)$ → 2nd level Gram matrix $O(mLN + m^2L)$. c.f. $O(m^2N^2)$

Big reduction if $L, n \ll N$

Comparison: Persistence Scale Space Kernel

(Reininghaus et al 2015)

- PSS Kernel

$$k_R(x,y) = \frac{1}{8\pi t}\left\{\exp\left(\frac{\|x-y\|^2}{8t}\right) - \exp\left(\frac{\|x-\bar{y}\|^2}{8t}\right)\right\}$$

$\bar{y} = (d,b)$ for $y = (b,d)$.

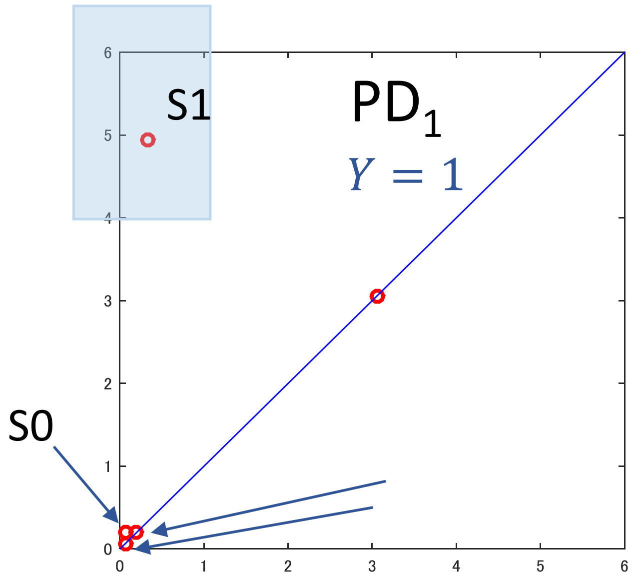
Pos. def. on $\{(b,d) | d \geq b\}$
0 on Δ.

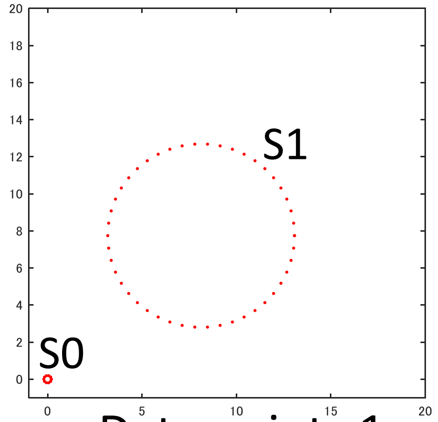
$\mathcal{E}_k(D)$ is considered.

- Comparison between PWGK and PSSK
  - PWGK can control the discount around the diagonal independently of the bandwidth parameter.
  - PSSK is not shift-invariant → Random feature approximation is not applicable.
  - In Reininghaus et al 2015, 2nd level kernel is not considered.

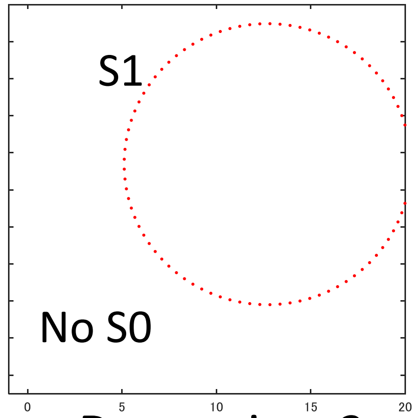
Synthetic example: SVM classification

- Classification of PD's by SVM
  - One big circle (random location and sample size) $S1$ with or without small circle $S0$.
  - $Y = \mathrm{XOR}(Z_1, Z_2)$
    - Z_1: Does S0 exists? Yes/No
    - Z_2: Is the generator of S1 within ((b($S1$)<1 && d($S1$))? Yes/No
  - Noise is added, in fact.
  - 100 for training and 100 for testing
  - Result (correct classification)
    - PWGK (proposed): 83.8%
    - PSSK (comparison): 46.5%





Data points 1



Data points 2

Applications

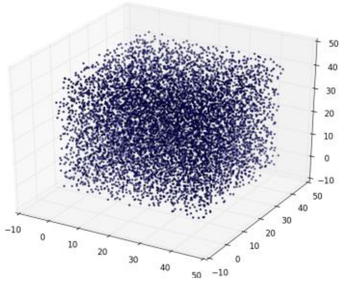
Application 1: Transition of Silica (SiO_2)

If cooled down rapidly from the liquid state, SiO_2 changes into the glass state (not to crystal).

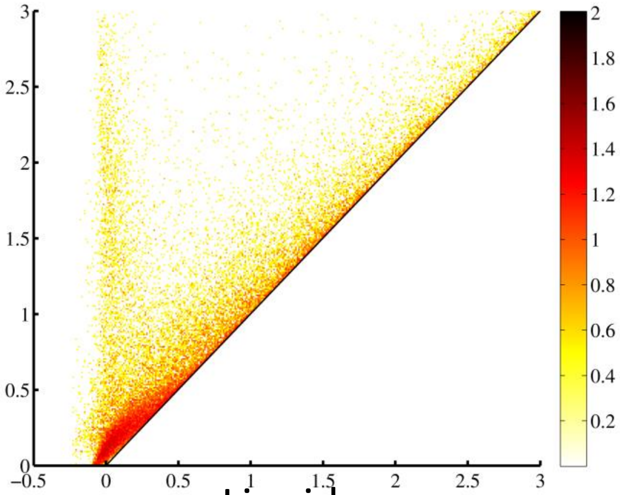
Amorphous: "soft" structure

Goal: identify the temperature of phase transition.

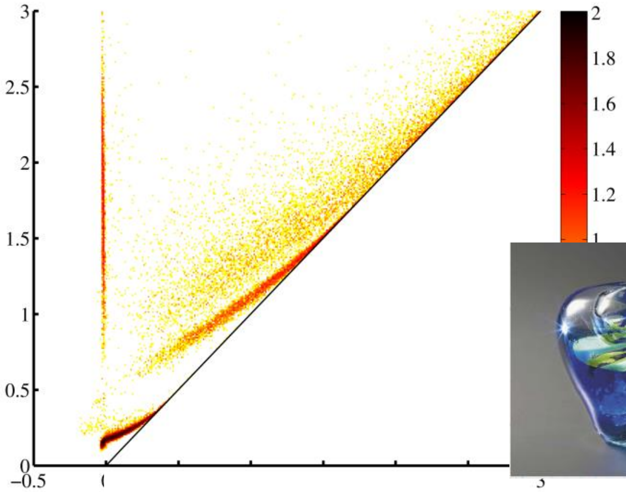
Data: Molecular Dynamics simulation for SiO_2. 3D arrangements of the atoms are used for computing PD at 80 temperatures. (Nakamura et al 2015; Hiraoka et al 2015)



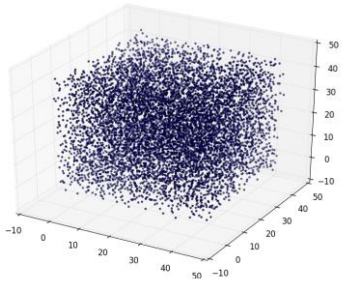
Examples of PD's

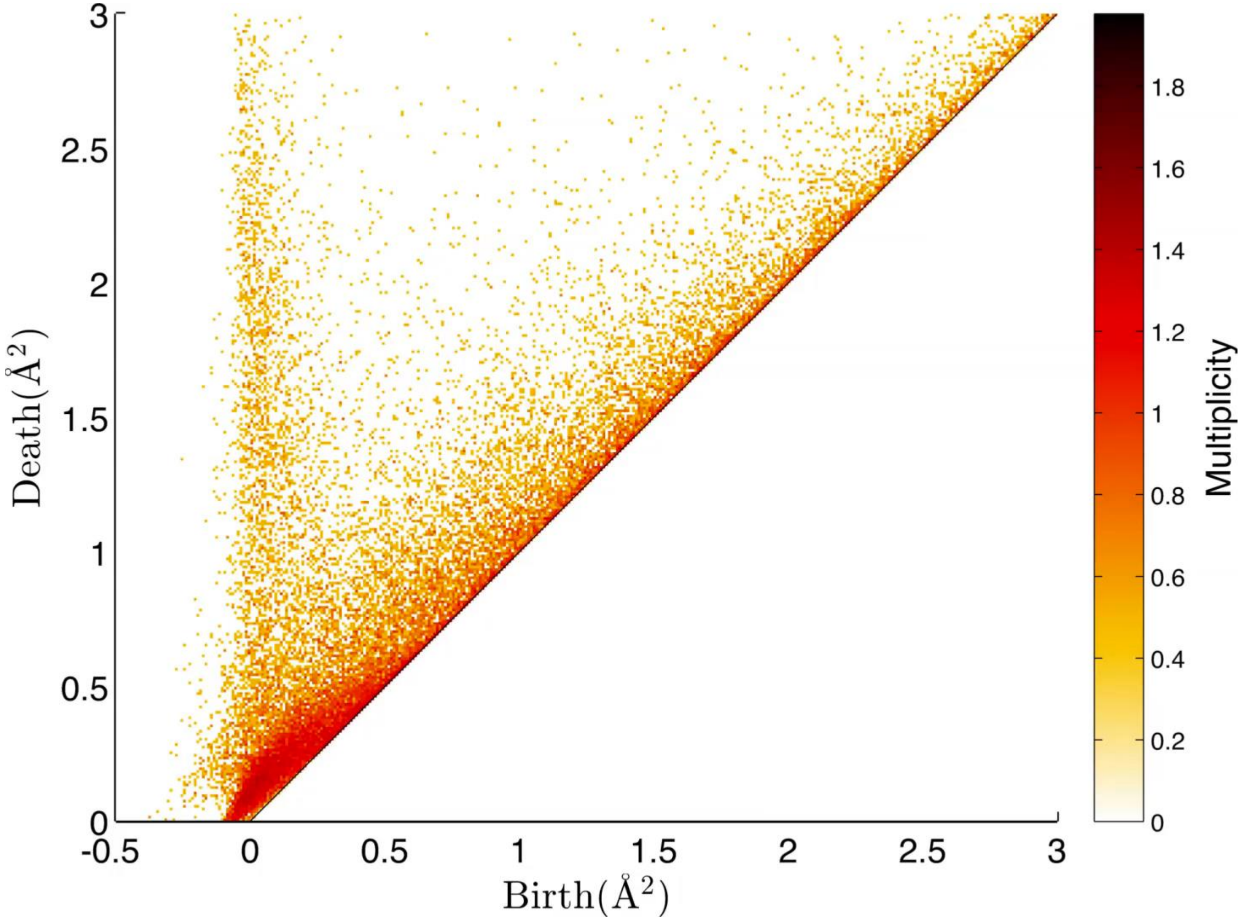


Liquid

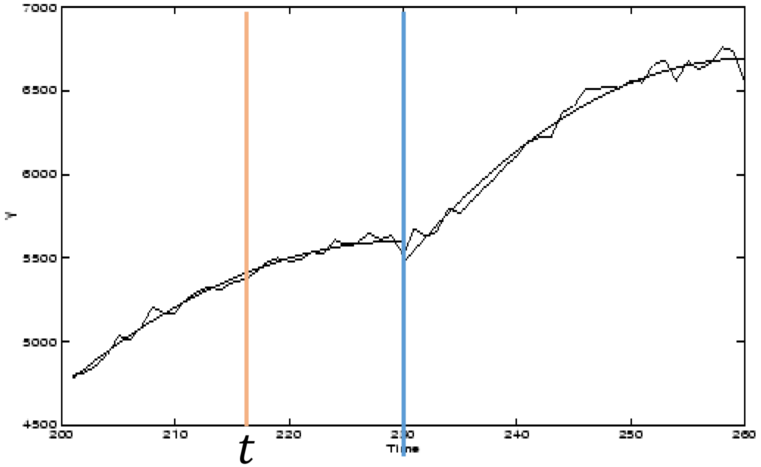


Glass (Amorphous)




Death(Å²)
Birth(Å²)
Multiplicity
3
2.5
2
1.5
1
0.5
0
-0.5
0
0.5
1
1.5
2
2.5
3
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0

Change point detection



- Data along a parameter t

  $X_t, t = 1, \dots, T.$

  Kernel Change Point Analysis with Fisher Discriminant score (Harchoui et al 2009):

  For each t, two classes are defined by the data before and after t.
  Fisher score on RKHS is used.

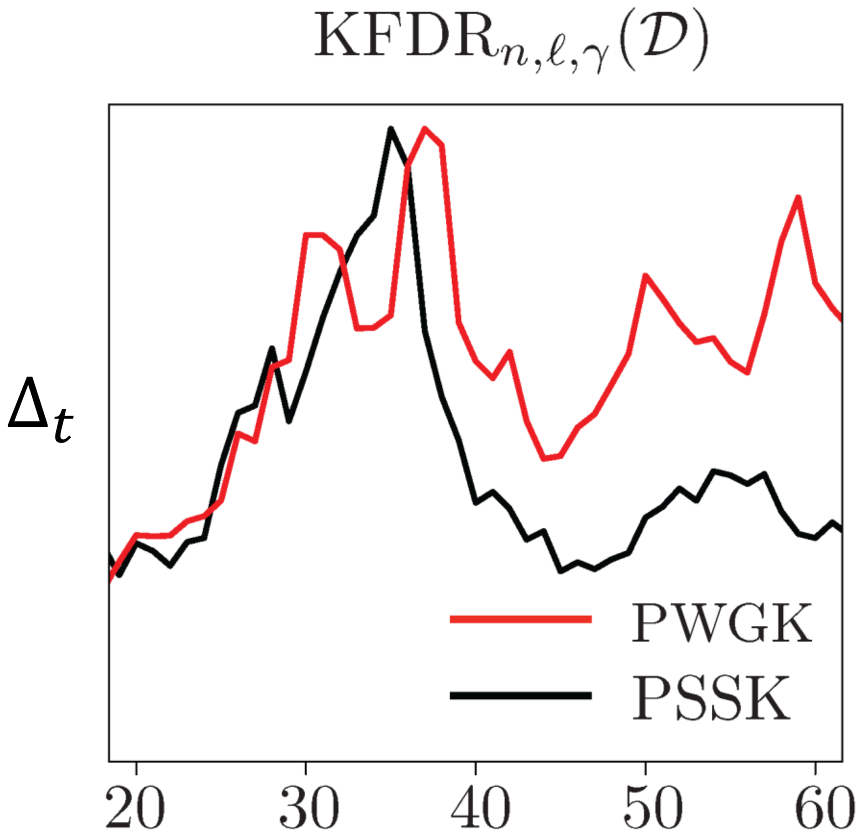
  - For each t, compute $\widehat{m}_{1:t} = \frac{1}{t}\sum_{i=1}^{t} \Phi(X_i)$ and $\widehat{m}_{t+1:T} = \frac{1}{T-t}\sum_{i=t+1}^{T} \Phi(X_i)$.
  - Compute $\Delta_t \coloneqq \left\| (V_{1:t} + V_{t+1:T} + \gamma I)^{-\frac{1}{2}} (\widehat{m}_{1:t} - \widehat{m}_{t+1:T}) \right\|_{H_k}^2$.
  - Find $\max_t \Delta_t$.

- For the packing problem, $X_t = \mathcal{E}_k\left(D_{\phi_t}\right)$ $(t = 1, \dots, 80)$.

- Detection of liquid-glass state transition
  - Approach in physics:

    Estimation using derivatives of enthalpy curve, but not so accurate.
  - Our approach: purely data-driven

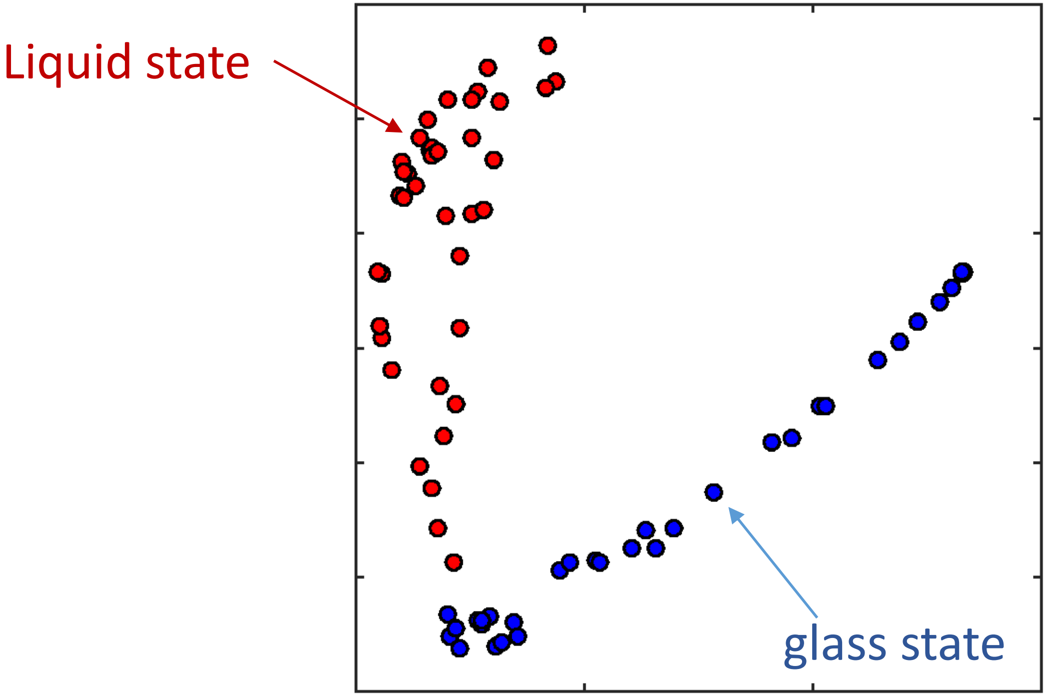
    Persistence diagrams, and then change point detection by Kernel FDR.
  - Number of generators in a PD is 30000 at most → difficult to use PSSK directly
  - PWGK (proposed) is applied with random features.



Detected change point = 3100K

Enthalpy by physicist: [2000K, 3500K]

- 2-dim plot by Kernel PCA



Sharp change between the two phases.

(Colored by the result of change point detection.
Colors are not used for KPCA).

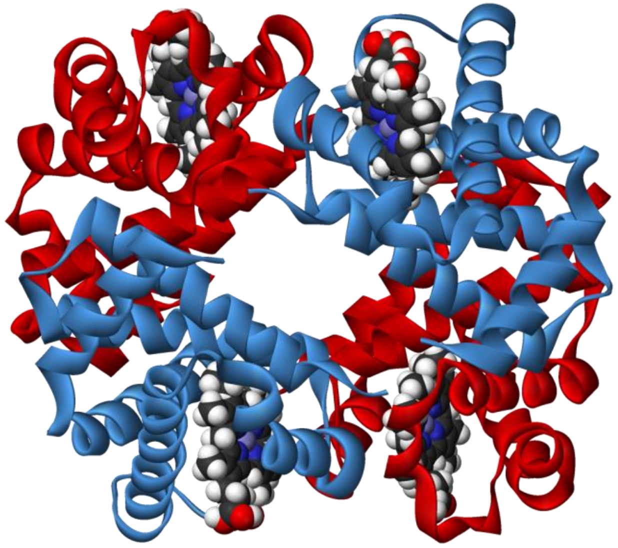
The result indicates that the phase can be identified by the snap-shot, while this is still controversial among physicists.

Application 2: Protein classification

- Structure of proteins → Functions
- The geometrical structure can be represented by persistence homology

- Classification of proteins with PD's.

  SVM is used.



- Data A: Protein-drug binding

- M2 channel in the influenza A virus:
  a target of medicine.
  Biding an inhibitor changes the structure

Cang, Mu, Wu, Opron, Xia, Wei, *Molecular Based Mathematical Biology* (2015) Fig. 3

- Task: Determine from the structure if there is rimantadine (inhibitor) in the M2 channel.

- Data: 3D-structures from NMR
  - 15 data for each of binding / non-binding.
  - Random choice of 10 training samples for each class. The rest is used for testing. 100 random choices for CV.

- Data B: 2 states of hemoglobin
  - Task: classify of the 2 states Relaxed (R) / Taut (T)
  - Data: 3D-stturcures from X-ray diffraction
    - R: 9 data, T: 10 data
    - Choice of one data from each class for testing, and the rest used for training.
    - All combinations are used for CV.

Relaxed (R)   Taut (T)

Cang, Mu, Wu, Opron, Xia, Wei, *Molecular Based Mathematical Biology* (2015) Fig. 4

- Results
  - Comparison with Cang et al (2015), where PH is used with 13 dimensional hand-made Molecular Topological Fingerprint (MTF) .
  - PWGK + SVM: only 1st PH is used.

MTF

#	Dim	Description
1	0	2nd longest lifetime
2	0	3rd longest lifetime
3	0	Total sum of lifetme
4	0	Average lifetime
5	1	Birth point of the longest generator
6	1	Longest lifetime
7	1	Birth points of the shortest generator among lifetime ≥1.5Å
8	1	Ave. medium points of generators among lifetime ≥1.5Å
9	1	Number of generators in [4.5, 5.5]Å, divided by total #atoms.
10	1	Number of generators in [3.5, 4.5)Å and (5.5, 6.5]Å, divided by total #atoms.
11	1	Total sum of lifetmes
12	1	Average lifetime
13	2	The birth point of the first generator.

CV classification rates

	A. Protein-Drug	B. Hemoglobin
PWGK	100	88.90
MTF*	(nbd) 93.91 / (bd) 98.31	84.50

* Results of MTF are taken from Cang et al. *Molecular Based Mathematical Biology* (2015).

Conclusion

- Topological data analysis
  - Key technology = persistence homology
  - PH can introduce useful features / descriptors for complex geometrical structures.
  - PH contains information more than topology.

- Statistical approach to topological data analysis
  - Statistical data analysis on many persistence diagrams.
  - Kernel methods introduce systematic data analysis to TDA.
    - Vectorization of persistence diagrams by kernel embedding.
    - Persistence weighted Gaussian kernel → flexible kernel for noise.

References

Kusano, G., Fukumizu, K., Hiraoka, Y. (2015) Persistence weighted Gaussian kernel for topological data analysis. *Proc. Intern. Conf. Machine Learning* 2016

Carlsson, G. (2009) Topology and data. *Bull. Amer. Math. Soc*., 46(2):255–308. http://dx.doi.org/10.1090/S0273-0979-09-01249-X .

Hiraoka, Y., Nakamura, T., Hirata, A., Escolar, E. G., Matsue, K., and Nishiura, Y. (2016) Description of medium-range order in amorphous structures by persistent homology. *PNAS*, 113(26), 7035–7040.

Nakamura, T., Hiraoka, Y., Hirata, A., Escolar, E. G., and Nishiura, Y. (2015) Persistent homology and many-body atomic structure for medium-range order in the glass. *Nanotechnology*, 26 (304001).

Reininghaus, J., Huber, S., Bauer, U., and Kwitt, R. (2015) A stable multi-scale kernel for topological machine learning. *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pp. 4741–4748.

Kwitt, R., Huber, S., Niethammer, M., Lin, W., and Bauer, U. (2015) Statistical topological data analysis - a kernel perspective. *Advances in Neural Information Processing Systems 28*, pp. 3052–3060.

Fasy, B. T., Lecci, F., Rinaldo, A., Wasserman, L., Balakrishnan, S., and Singh, A. (2014) Confidence sets for persistence diagrams. *The Annals of Statistics*, 42(6):2301–2339,

Cang, Z., Mu, L., Wu, K., Opron, K., Xia, K., and Wei, G. W. (2015) A topological approach for protein classification. *Molecular Based Mathematical Biology*, 3(1), 2015