Introduction to topological data analysis

Ippei Obayashi

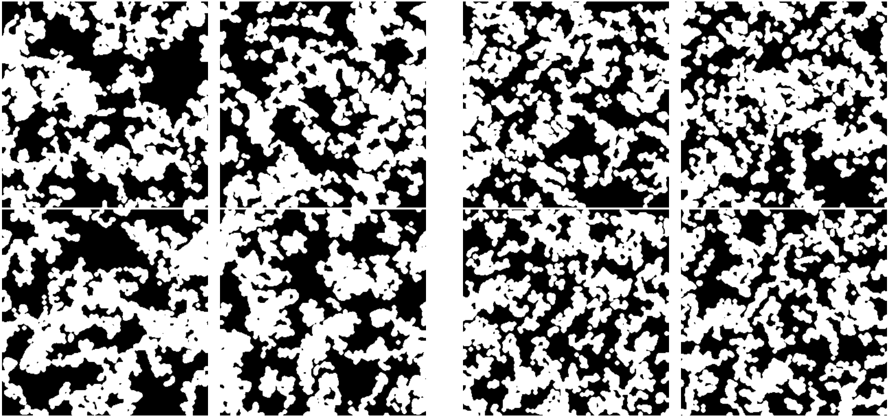
Adavnced Institute for Materials Research, Tohoku University

Jan. 12, 2018

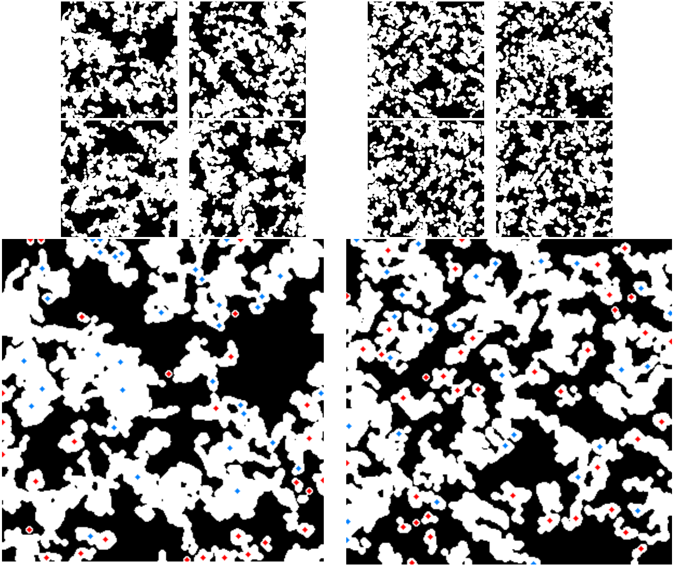
Persistent homology

- Topological Data Analysis (TDA)
  - Data analysis methods using topology from mathematics
  - Characterize the shape of data quantitatively
    - By using connected components, rings, cavities, etc.
- Persistent homology (PH) is a main tool of TDA
  - The key idea is "Homology" from mathematics
  - Gives a good descriptor for the shape of data (called a persistence diagram)
- Rapidly developed in 21st century
  - Mathematical theories
  - Software
  - Applications to materials science, sensor network, phylogenetic network, etc.

Example 1

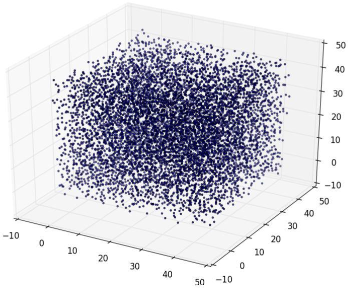


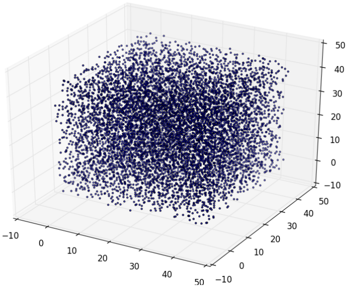
These images are classified into two groups (left 4 images and right 4 images). Do you find the characteristic shape to distinguish the two groups?



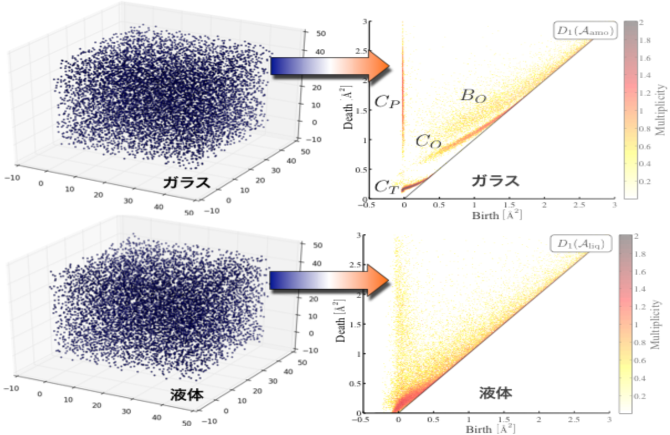
Shapes around blue dots are “typical” for left images, and red dots for right images

Example 2





Atomic configurations of amorphous silica (SiO_2) and liquid silica. Do you find the difference?

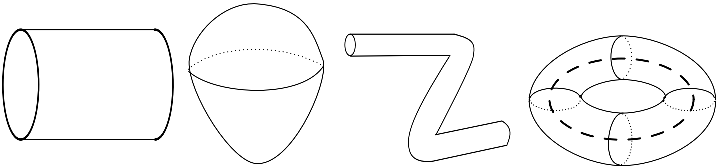


From Y. Hiraoka, et al., PNAS 113(26):7035-40 (2016)

Persistence diagrams can capture the difference clearly

Homology

- Connected components, rings, and cavities are mathematically formalized by homology.
- Algebra is used to formalize such geometric structures
- There are many types of holes and characterized by "dimension"

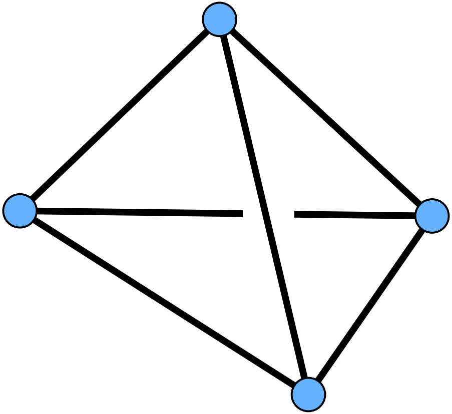


dim 1: 1	dim 1: 0	dim 1: 1	dim 1: 2
dim 2: 0	dim 2: 1	dim 2: 0	dim 2: 1

1 dim: You can see the inside from outside   2 dim: You cannot see

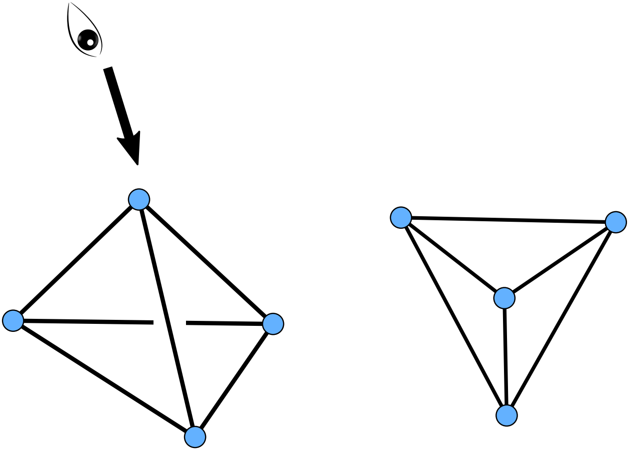
How to count rings

How many rings/holes in the tetrahedron skelton?

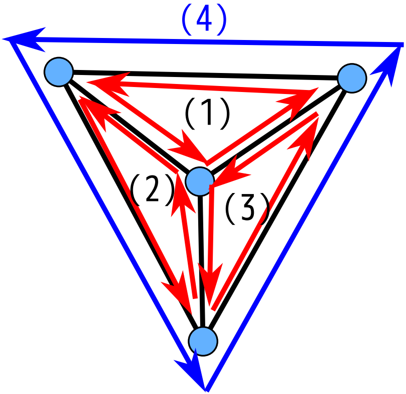


Four?

But if you see the tetrahedron from upside, the number of rings is three.



What happened?



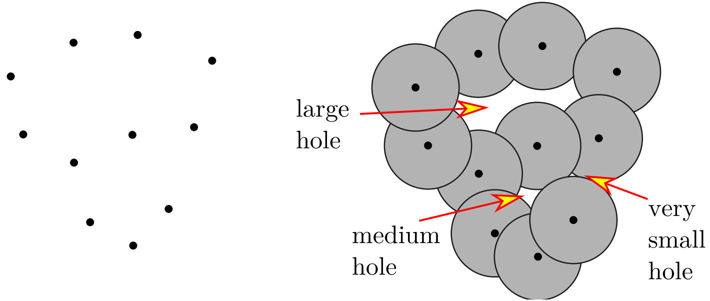
We cosider the addition of rings. Then $(1) + (2) + (3) = (4)$ since two arrows with opposite directions are vanished when added. This means that the four rings are not *linearly independent*. We can formalize the number of linearly independent rings by linear algebra.

Persistent homology

- Characterizing the shape of data is a difficult problem
  - Especially, for 3D data
- Homology is one possible tool for that purpose, but homology drops the details about the shape of data too much
  - Homology can only count the number of holes
- We want more information about the shape of data with easy-to-use form
- Computational homology is proposed in 20 century, but it is sensitive to noise

$\rightarrow$ using increasing sequence (called filtration)

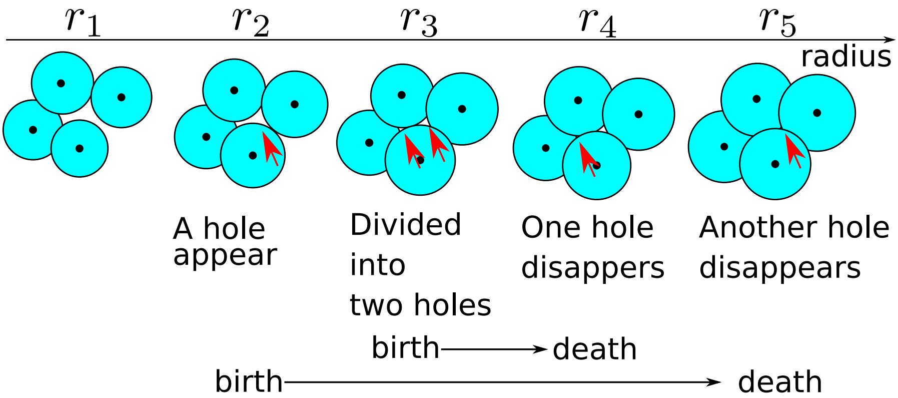
r-Ball model



- Input data is a set of points (called a point cloud)
- The points themselves have no "hole", but there are some hole-like structures
- Put a disc whose radius is r onto each point
- There are three holes
  - Homology can detect the number of holes

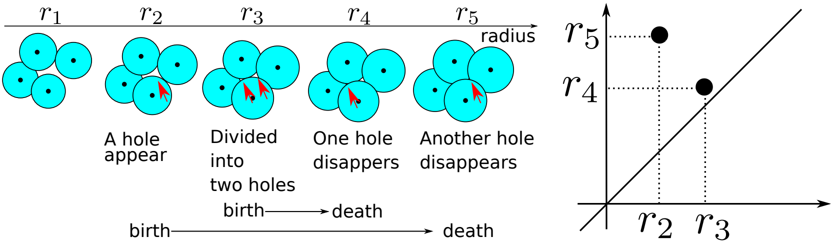
Filtration

By increasing the radii r gradually, many holes appear and disappear. The theory of PH can make mathematically proper pairs of the radii of appearance and disappearance.



Persistence diagram

The pairs are called birth-death pairs. The pairs are visualized by a scatter plot on (x, y)-plane.



This diagram visualizes 1-dimensional persistent homology. This diagram is called persistence diagram.

- We can apply PH to any dimensional data.
  - Practical for 2D and 3D
  - Because it is difficult to understand high dimensional "holes"
  - Since it is hard to characterize the shape of 3D data, the application to 3D data is especially useful
- We can apply PH to various kinds of increasing sequences
  - We can apply PH other than point clouds
  - Bitmap data
  - PH is useful for 3D bitmap data such as X-ray CT data

Mathematics of PH

PH relates various fields

- Algebraic topology
- Representation theory
- Computational geometry
- Combinatorics
- Probability theory
- Statistics

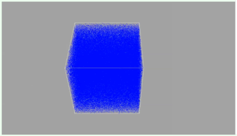
Various studies about fundamental theories are important

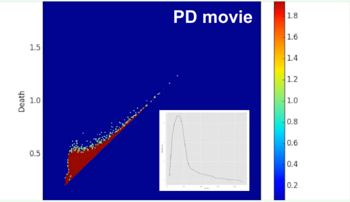
Ichinomiya, Obayashi, H. PRE (2017)

Craze formation of polymers

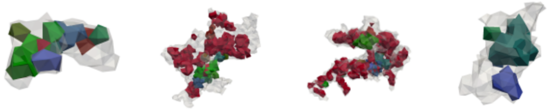
Kremer-Grest model

uniaxial deformation



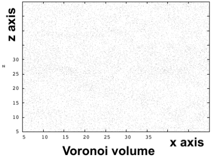


void coalescence during craze formation

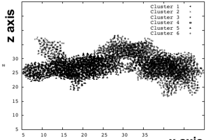


- gray voids are large voids observed after yielding
- color voids are initial micro voids generating large voids

craze position



Voronoi volume (conventional)



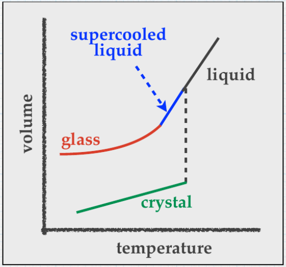
PD method

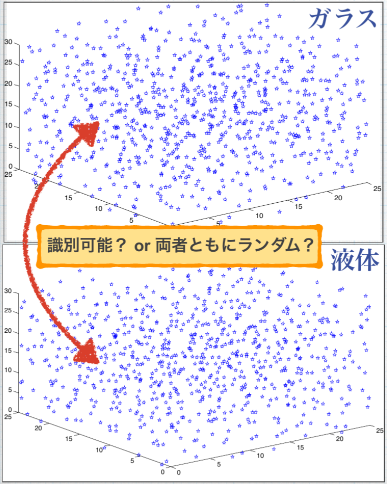
- detect large voids from PD movie by generators with large death values
- explore initial configurations of large voids by reversing time
- large voids are generated by coalesce of micro voids (void percolation)

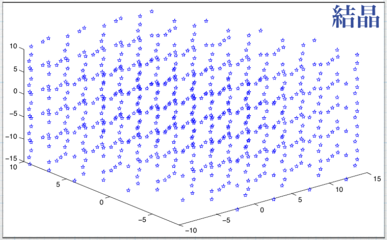
Amorphous Silica

- What is glass?
- Not liquid, not solid, but something in-between
- Atomic configuration looks random
- But it maintains rigidity
- We require further geometric understandindgs of atomic configurations

シリカの原子配置

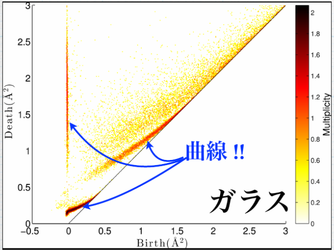

supercooled liquid
liquid
volume
glass
crystal
temperature

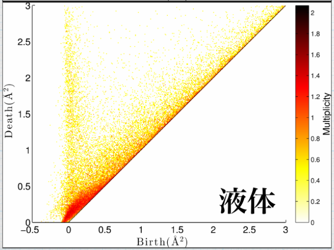

ガラス
識別可能？ or 両者ともにランダム？
液体


結晶

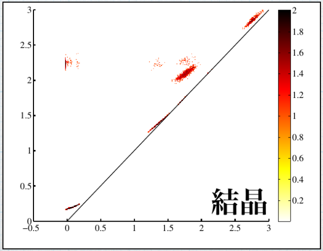
Y.H., T. Nakamura et al. PNAS (2016)

シリカのパーシステント図



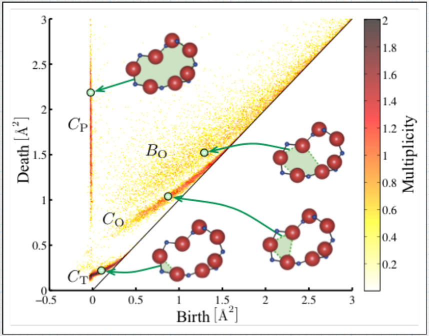


- PD1を表示（リング構造に着目）
- 結晶の規則性は 0次元的分布
- 液体のランダム性は 2次元的分布
- ガラスは 1次元的分布（曲線）!!



Y.H., T. Nakamura et al. PNAS (2016)

ガラスの階層的幾何構造



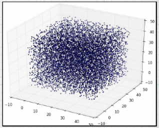
ガラスの幾何構造

⇕

PD内の曲線の幾何学的な起源

逆問題 →

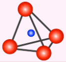
- optimal cycle
  Escolar and H. 2015.
- continuation
  Gameiro, Obayashi, H. Physica D, 2015



階層的リング構造

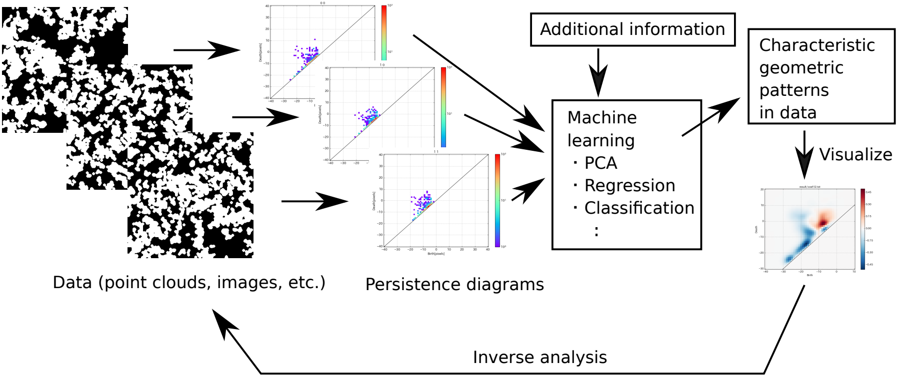
C_P: primary rings generating the others → **C_O: three oxygen rings**

↙ **C_T: triangles on tetrahedra**



↘ **B_O: oxygen rings ($\geq$ four)**

Combination of statistics/machine learning



Software

For the practical data analysis using PH, analysis software is important.
I will introduce Homcloud.

Softwares for PH

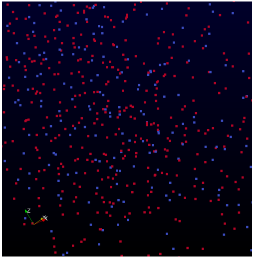
Various analysis softwares are developed for their own purpose and interest

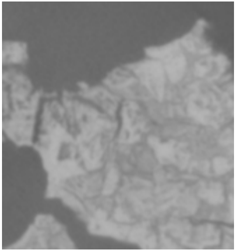
- Gudhi
- dipha, phat, ripser
- eirine
- RIVET
- JavaPlex
- Perseus
- Dionysus

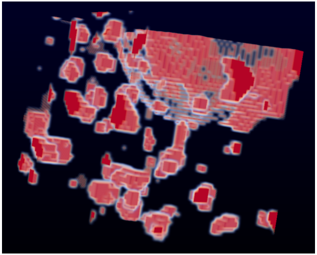
⋮

Homcloud

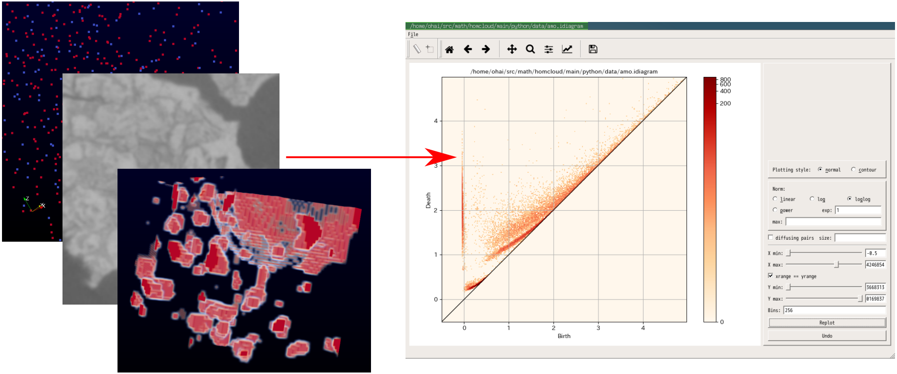
- Focus on applications, especially to materials science
  - Data analysis for molecular dynamical simulations
  - Images from electric microscopy, 3D images from X-ray CT



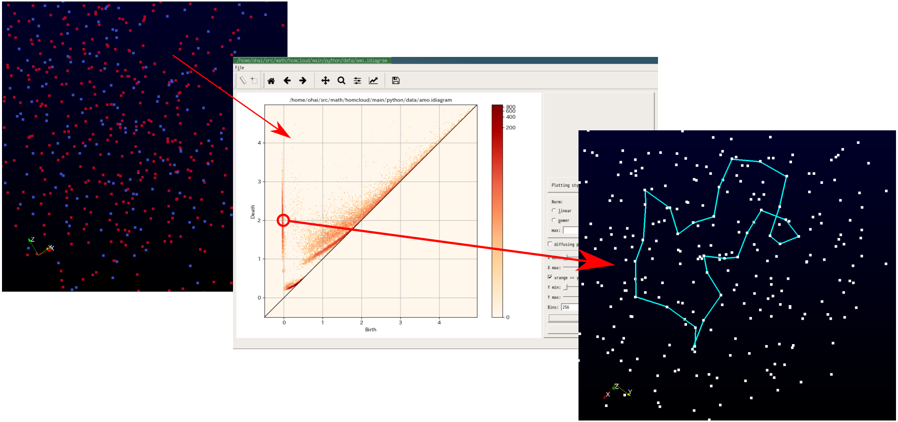




We can compute persistence diagrams from various sources (point clouds, 2D/3D bitmap data)



Inverse analysis



Homcloud as a platform for the development of new methods

- Getting an idea → Writing a code and trying it → If it works, we consider a background theory
- We can quickly introduce such a new idea into data analysis
  - Collaborators also use the idea quickly
- Try ideas found in papers by other researchers

- I develop the software and analyze data together
  - Mainly data from materials science
    - Provided by collaborators
  - Dogfooding
  - Do not implement unused functionality
  -
- Collaborators also use Homcloud
- Implemented mainly in python
  - Python is often used for data science

Homcloud Demo

Future plan of Homcloud

- Better user interface
- Performance improvement
- Implement new methods
  - Parallel to theoretical researches
- Publish in this winter
  - http://www.wpi-aimr.tohoku.ac.jp/hiraoka_labo/homcloud.html
- If you want to use Homcloud, please contact with us: ippei.obayashi.d8@tohoku.ac.jp

Wrap up

- Persistent homology enable us to analyze the shape of data quantitatively and effectively by using the power of the mathematical theory of topology
  - A persistence diagram is a good descriptor for the shape of data
  - Applications to 3D data is most effective, in my opinion
- There are many applications
  - We mainly apply persistent homology to materials science
  - Meteology
  - Brain science, life science, etc.
- Combination of theoretical researches, software development, and applications is important