EMAp Summer Course

Topological Data Analysis with Persistent Homology

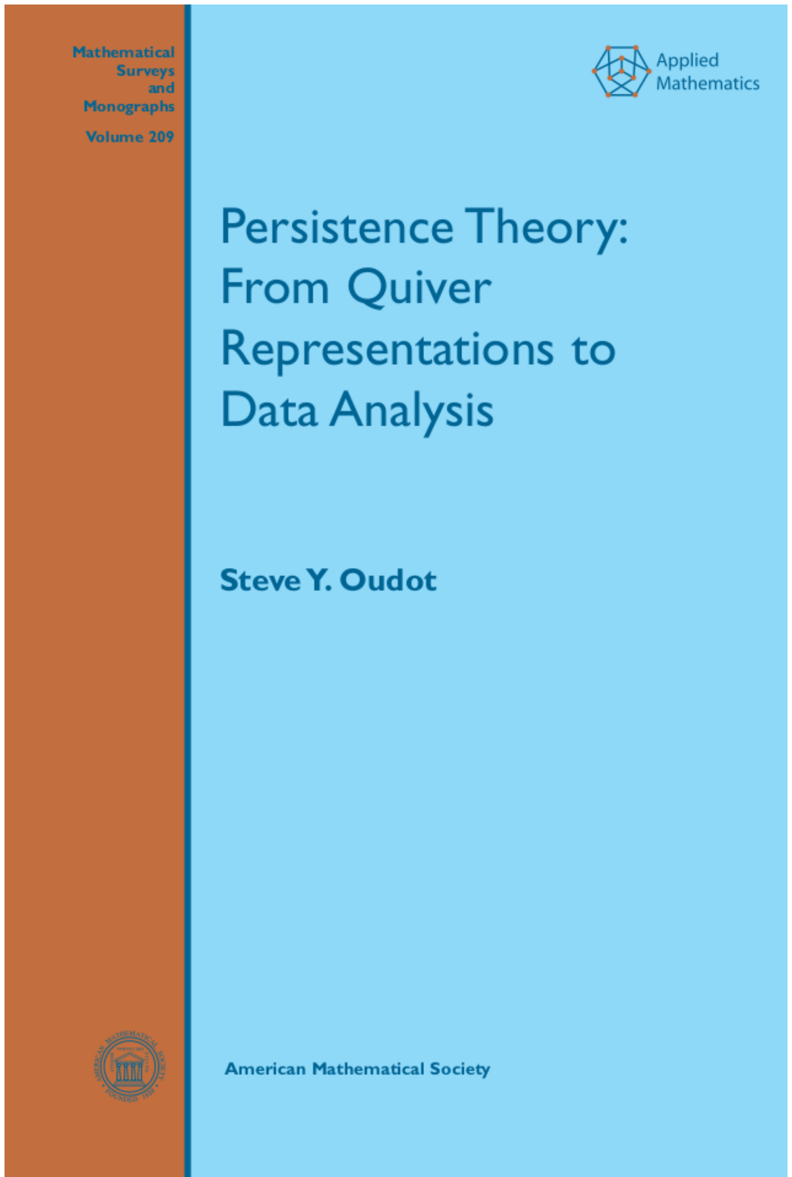
https://raphaeltinarrage.github.io/EMAp.html

Lesson 8: Datasets have topology

Last update: February 4, 2021

Introduction

Oudot in 2015



Applications. This richness is also reflected in the diversity of the applications, whose list has been ever growing since the early developments of the theory. The following excerpt[5] illustrates the variety of the topics addressed:

- analysis of random, modular and non-modular scale-free networks and networks with exponential connectivity distribution [158],
- analysis of social and spatial networks, including neurons, genes, online messages, air passengers, Twitter, face-to-face contact, co-authorship [210],
- coverage and hole detection in wireless sensor fields [98, 136],
- multiple hypothesis tracking on urban vehicular data [23],
- analysis of the statistics of high-contrast image patches [54],
- image segmentation [70, 209],
- 1d signal denoising [212],
- 3d shape classification [58],
- clustering of protein conformations [70],
- measurement of protein compressibility [135],
- classification of hepatic lesions [1],
- identification of breast cancer subtypes [205],
- analysis of activity patterns in the primary visual cortex [224],
- discrimination of electroencephalogram signals recorded before and during epileptic seizures [237],
- analysis of 2d cortical thickness data [82],
- statistical analysis of orthodontic data [134, 155],
- measurement of structural changes during lipid vesicle fusion [169],
- characterization of the frequency and scale of lateral gene transfer in pathogenic bacteria [125],
- pattern detection in gene expression data [105],
- study of plant root systems [115, §IX.4],
- study of the cosmic web and its filamentary structure [226, 227],
- analysis of force networks in granular matter [171],
- analysis of regimes in dynamical systems [25].

In most of these applications, the use of persistence resulted in the definition of new descriptors for the considered data, which revealed previously hidden structural information and allowed the authors to draw original conclusions.

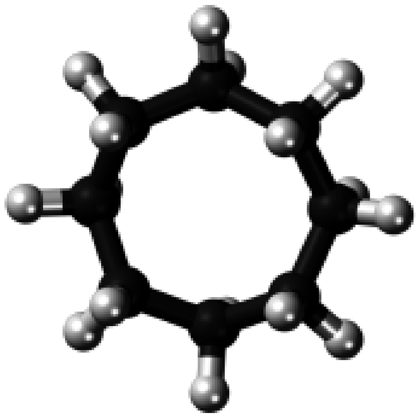
I - Some examples

II - Betti curves

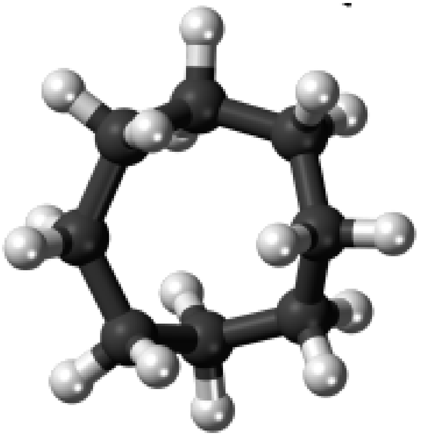
(III - Tutorial)

Shawn Martin, Aidan Thompson, Evangelos A Coutsias, and Jean- Paul Watson. *Topology of cyclo-octane energy landscape.* The journal of chemical physics, 2010. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3188624/

The cyclo-octane molecule C_8H_{16} admits several stable configurations, i.e., several spatial arrangements of its atoms.



crown

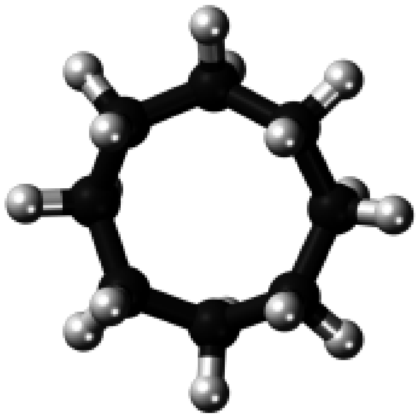


boat-chair

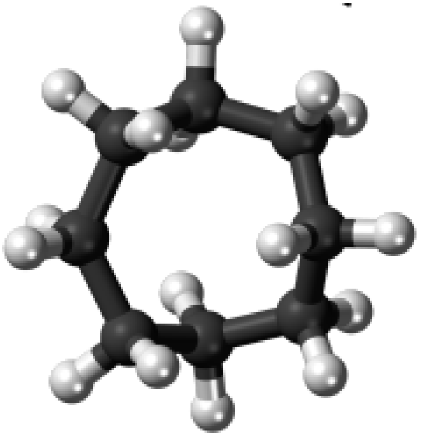
Configurations of cyclo-octane

Shawn Martin, Aidan Thompson, Evangelos A Coutsias, and Jean- Paul Watson. *Topology of cyclo-octane energy landscape.* The journal of chemical physics, 2010. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3188624/

The cyclo-octane molecule $\mathrm{C_8H_{16}}$ admits several stable configurations, i.e., several spatial arrangements of its atoms.

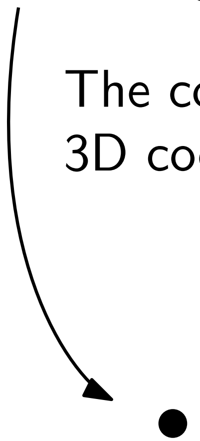


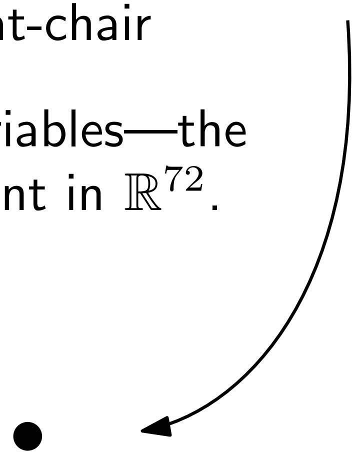
crown



boat-chair

The configuration of such a molecule can be represented by 72 variables—the 3D coordinates of each of its 24 atoms—, or equivalently, by a point in $\mathbb{R}^{72}$.

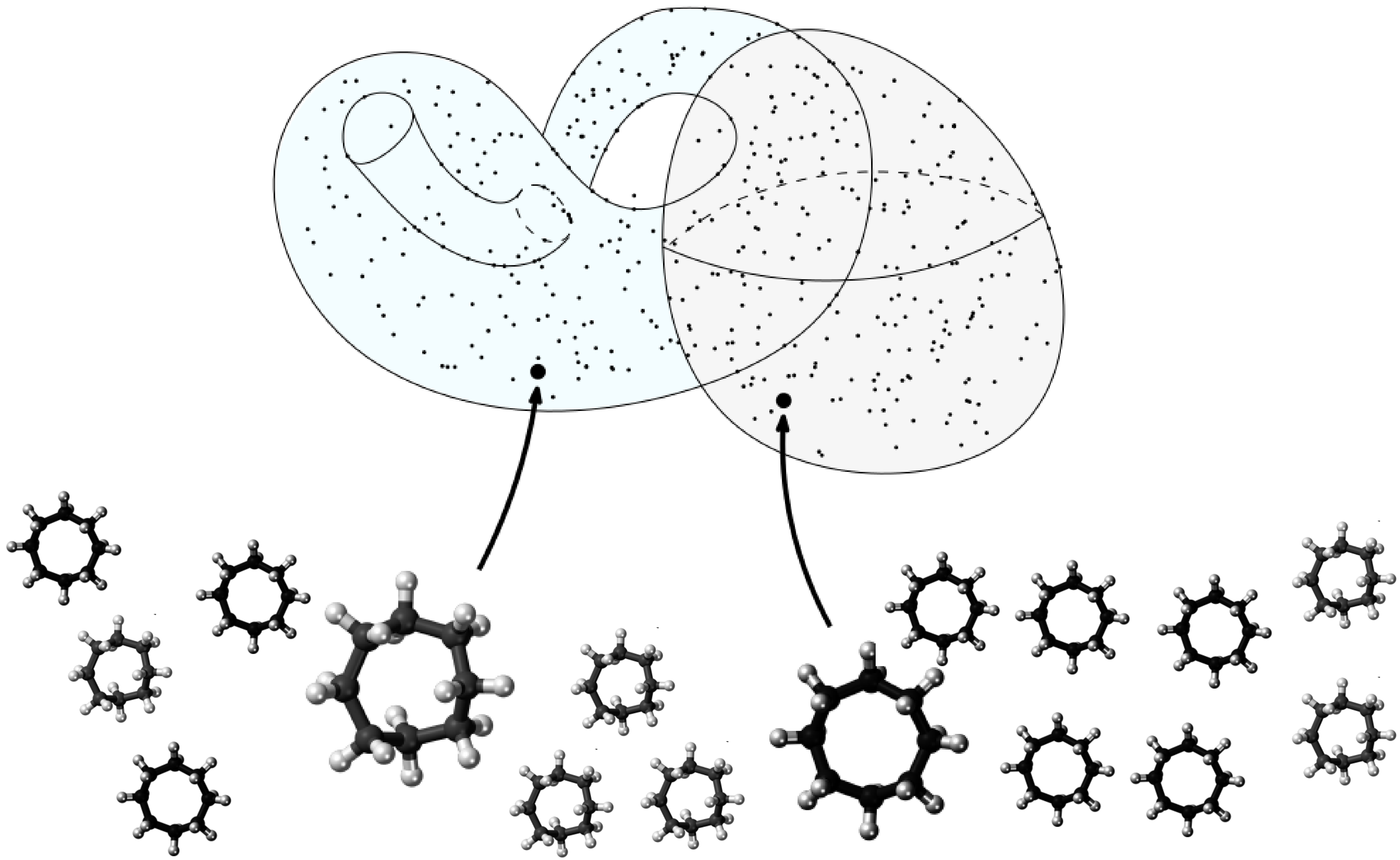




Configurations of cyclo-octane

The configuration of such a molecule can be represented by 72 variables—the 3D coordinates of each of its 24 atoms—, or equivalently, by a point in $\mathbb{R}^{72}$.

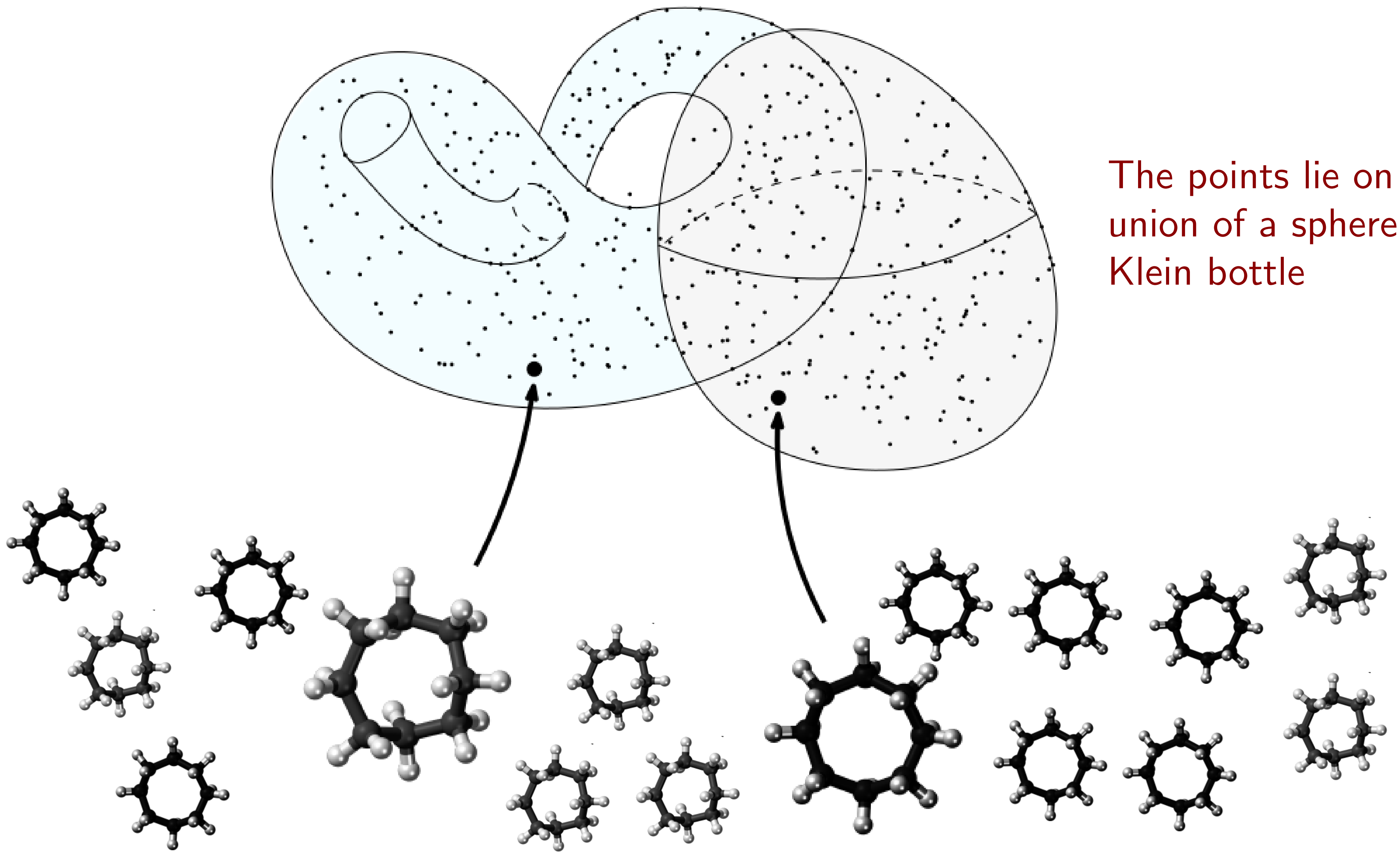
By analyzing many of these molecules, the authors obtain **a point cloud in** $\mathbb{R}^{72}$.



Configurations of cyclo-octane

The configuration of such a molecule can be represented by 72 variables—the 3D coordinates of each of its 24 atoms—, or equivalently, by a point in $\mathbb{R}^{72}$.

By analyzing many of these molecules, the authors obtain **a point cloud in $\mathbb{R}^{72}$**.

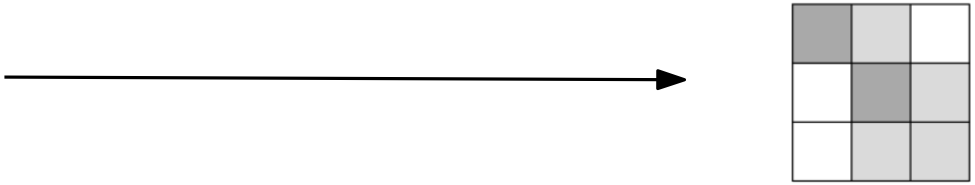


The points lie on the union of a sphere and a Klein bottle

Natural images

Gunnar Carlsson, Tigran Ishkhanov, Vin De Silva, and Afra Zomorodian. *On the local behavior of spaces of natural images.* International journal of computer vision, 2008. `https://www.researchgate.net/publication/220659347OntheLocalBehaviorofSpacesofNaturalImages`

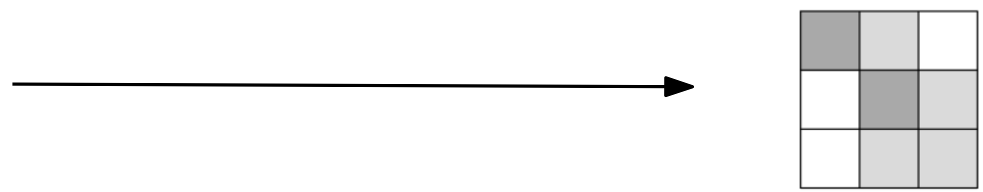
From a large collection of natural images, the authors extract 3×3 patches.



Natural images

Gunnar Carlsson, Tigran Ishkhanov, Vin De Silva, and Afra Zomorodian. *On the local behavior of spaces of natural images.* International journal of computer vision, 2008. `https://www.researchgate.net/publication/220659347OntheLocalBehaviorofSpacesofNaturalImages`

From a large collection of natural images, the authors extract 3×3 patches.

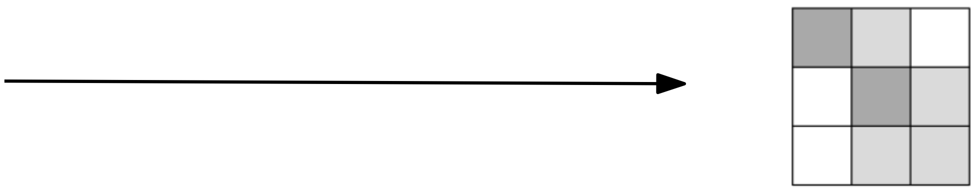


Since it consists of 9 pixels, each of these patches can be seen as point in $\mathbb{R}^9$, and the whole set as a point cloud in $\mathbb{R}^9$.

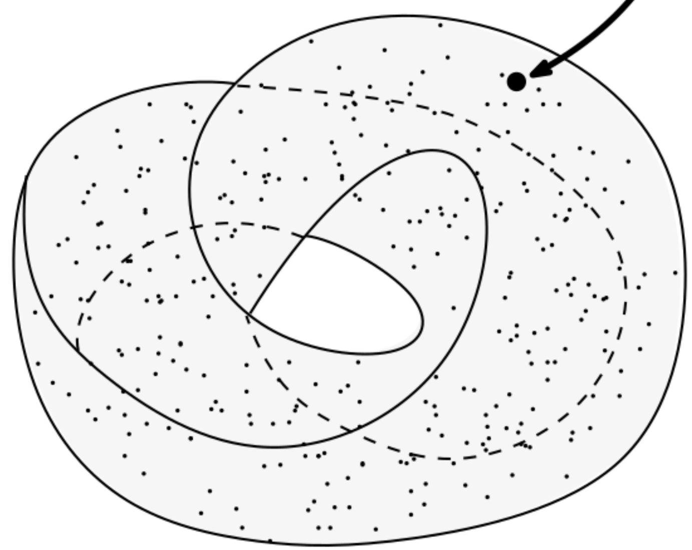
Natural images

Gunnar Carlsson, Tigran Ishkhanov, Vin De Silva, and Afra Zomorodian. *On the local behavior of spaces of natural images.* International journal of computer vision, 2008. `https://www.researchgate.net/publication/220659347OntheLocalBehaviorofSpacesofNaturalImages`

From a large collection of natural images, the authors extract 3×3 patches.



Since it consists of 9 pixels, each of these patches can be seen as point in $\mathbb{R}^9$, and the whole set as a point cloud in $\mathbb{R}^9$.



this dataset concentrates near a Klein bottle

Breast cancer

Monica Nicolau, Arnold J Levine, and Gunnar Carlsson. *Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival.* Proceedings of the National Academy of Sciences, 2011.
https://www.pnas.org/content/108/17/7265

Tissues of patients infected with breast cancer has been analyzed, resulting in 262 genomic variables per patients.

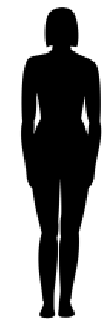
$(x_1, x_2, ..., x_{262})$

Breast cancer

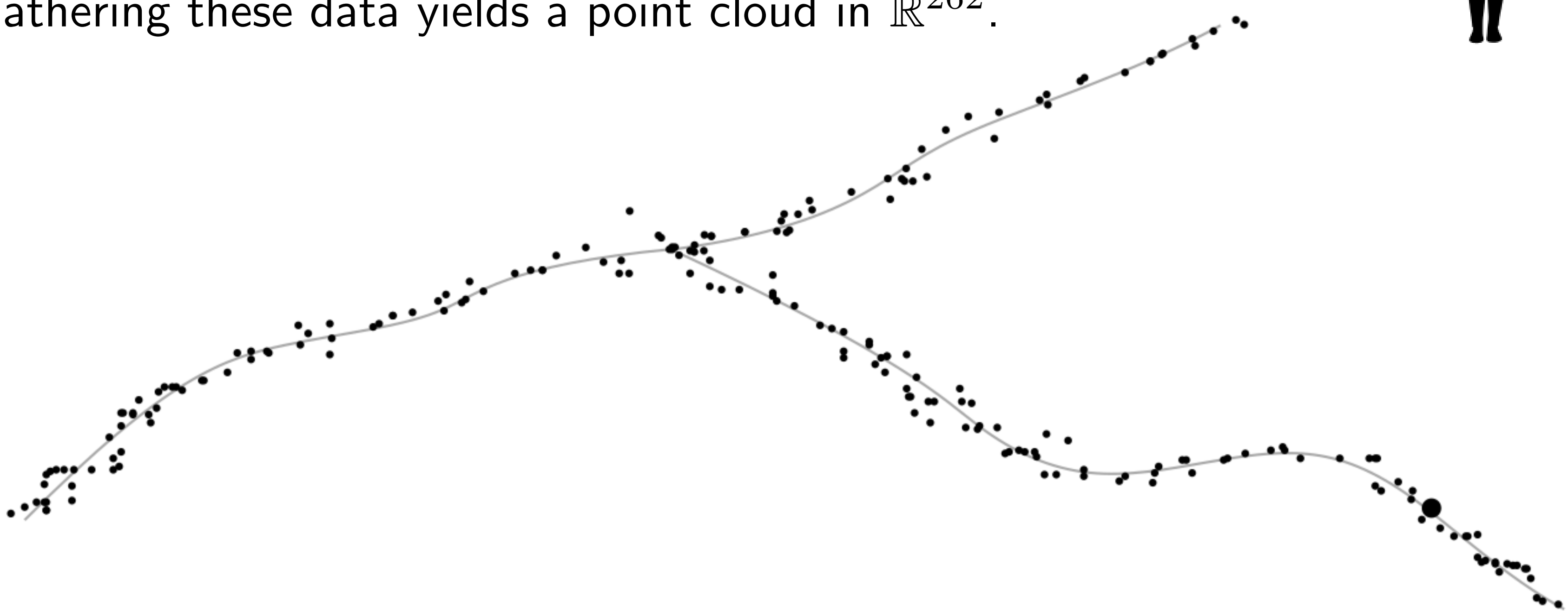
Monica Nicolau, Arnold J Levine, and Gunnar Carlsson. *Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival.* Proceedings of the National Academy of Sciences, 2011.
`https://www.pnas.org/content/108/17/7265`

Tissues of patients infected with breast cancer has been analyzed, resulting in 262 genomic variables per patients.

$$(x_1, x_2, \ldots, x_{262}) \longleftarrow$$



Gathering these data yields a point cloud in $\mathbb{R}^{262}$.



Breast cancer

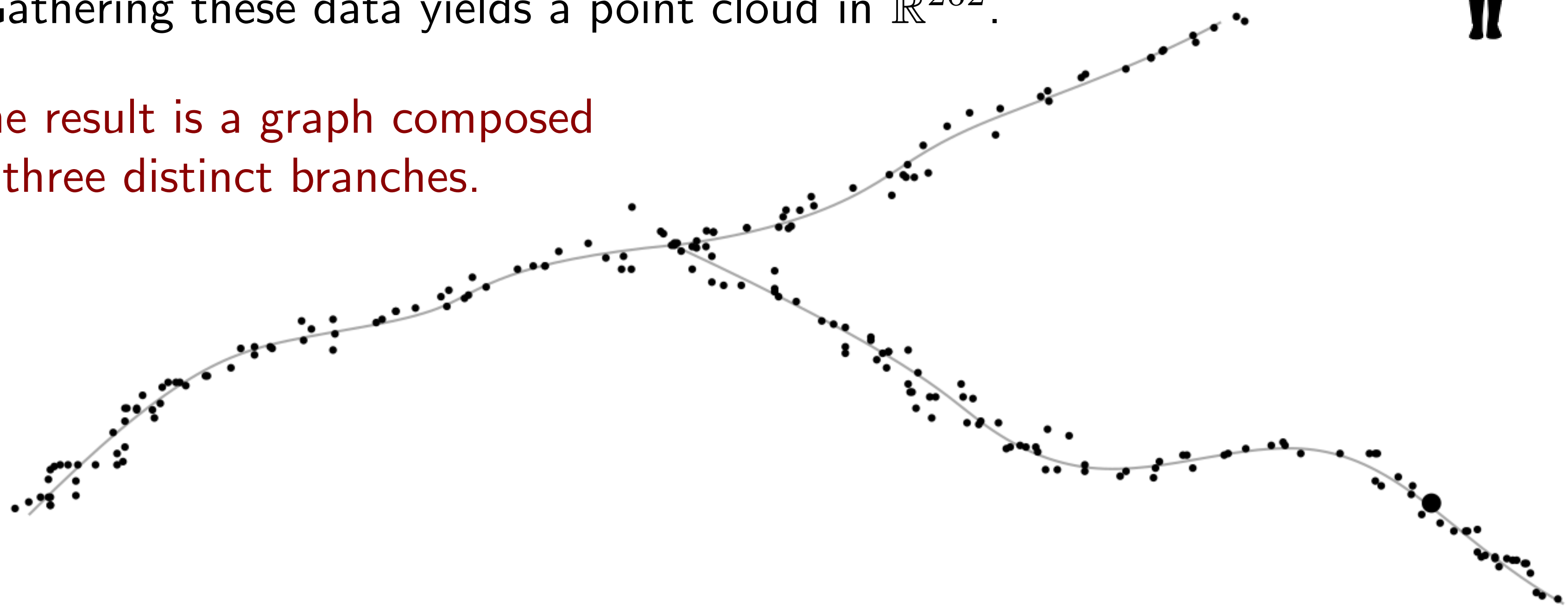
Monica Nicolau, Arnold J Levine, and Gunnar Carlsson. *Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival.* Proceedings of the National Academy of Sciences, 2011.
`https://www.pnas.org/content/108/17/7265`

Tissues of patients infected with breast cancer has been analyzed, resulting in 262 genomic variables per patients.

$$(x_1, x_2, \ldots, x_{262}) \longleftarrow$$

Gathering these data yields a point cloud in $\mathbb{R}^{262}$.

The result is a graph composed of three distinct branches.

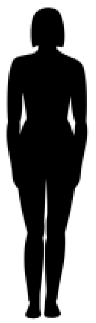


Breast cancer

Monica Nicolau, Arnold J Levine, and Gunnar Carlsson. *Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival.* Proceedings of the National Academy of Sciences, 2011.
`https://www.pnas.org/content/108/17/7265`

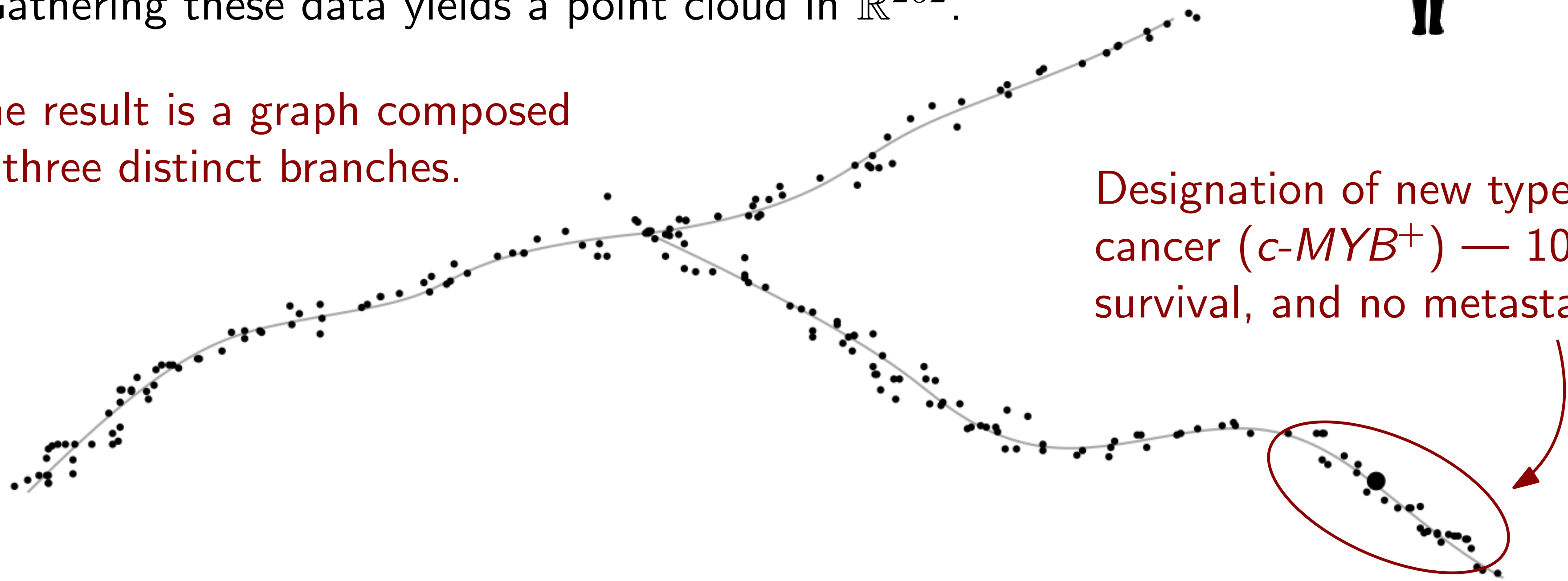
Tissues of patients infected with breast cancer has been analyzed, resulting in 262 genomic variables per patients.

$(x_1, x_2, ..., x_{262})$ ⟵



Gathering these data yields a point cloud in $\mathbb{R}^{262}$.

The result is a graph composed of three distinct branches.



Designation of new type of breast cancer (*c-MYB*$^+$) — 100% survival, and no metastasis.

I - Some examples

II - Betti curves

(III - Tutorial)

Reminder of yesterday

Let $\mathcal{M} \subset \mathbb{R}^n$ be a bounded subset.
Suppose that we are given a finite sample $X \subset \mathcal{M}$.
Estimate the homology groups of $\mathcal{M}$ from X.

Definition: For every $t \geq 0$, the t-*thickening* of the set X, denoted X^t, is the set of points of the ambient space with distance at most t from X:

$$X^t = \bigcup_{x \in X} \overline{\mathcal{B}}\left(x, t\right).$$

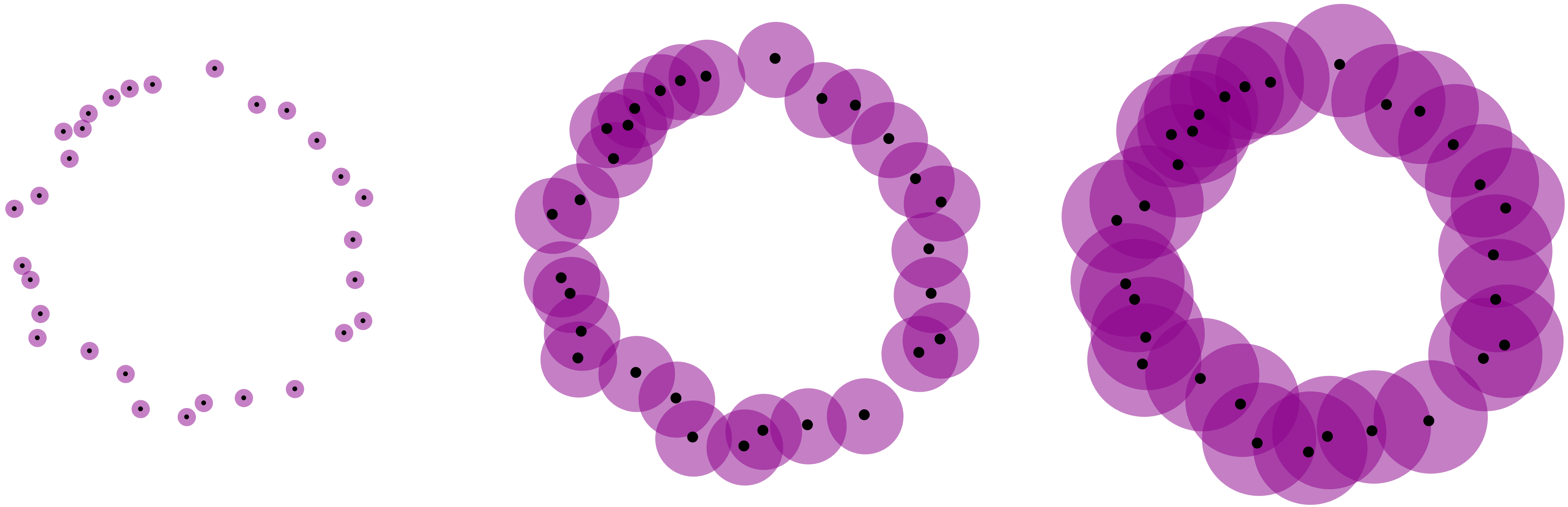
Definition: The *Čech complex of X at time t* is the nerve of the cover

$$\mathcal{V}^t = \left\{\overline{\mathcal{B}}\left(x, t\right), x \in X\right\}.$$

Definition: The *Rips complex of X at time t* is the clique complex of the graph G^t defined as: its vertex set is $\{1, \ldots, N\}$, and its edges are the pairs (i, j) such that $\|x_i - x_j\| \leq 2t$.

Betti numbers

We can compute the Betti numbers for each value of t:



$X^{0.05}$ | $X^{0.2}$ | $X^{0.3}$

$\beta_0 = 30$, $\beta_1 = 0$ | $\beta_0 = 1$, $\beta_1 = 0$ | $\beta_0 = 1$, $\beta_1 = 1$

Betti curves

Definition: Let $X \subset \mathbb{R}^n$ and $i \geq 0$. The i^{th} *Betti curve* of X is the map

$$\begin{aligned} \beta_i(t) \colon \mathbb{R}^+ &\longrightarrow \mathbb{N} \\ t &\longmapsto \beta_i(X^t) \end{aligned}$$

In our context, this will be

$$\begin{aligned} \beta_i(t) \colon \mathbb{R}^+ &\longrightarrow \mathbb{N} \\ t &\longmapsto \beta_i\left(\mathrm{Rips}^t(X)\right) \end{aligned}$$

Exercise: For $i = 0$, show that $t \mapsto \beta_0(t)$ is non-increasing.

I - Some examples

II - Betti curves

(III - Tutorial)

Link of the notebook

https://github.com/raphaeltinarrage/EMAp/blob/main/Tutorial2.ipynb

Conclusion

We tried to find topology in datasets.

We studied it via the Betti curves.

We are ready for Persistent Homology :)

Homework: não

Conclusion

We tried to find topology in datasets.

We studied it via the Betti curves.

We are ready for Persistent Homology :)

Homework: não

Merci !