Schema Inference

Algebraic Topology for Schema Discovery

Luciano Melodia

Chair of Computer Science 6
Friedrich-Alexander University

luciano.melodia@fau.de

July 17, 2019

FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG

Overview

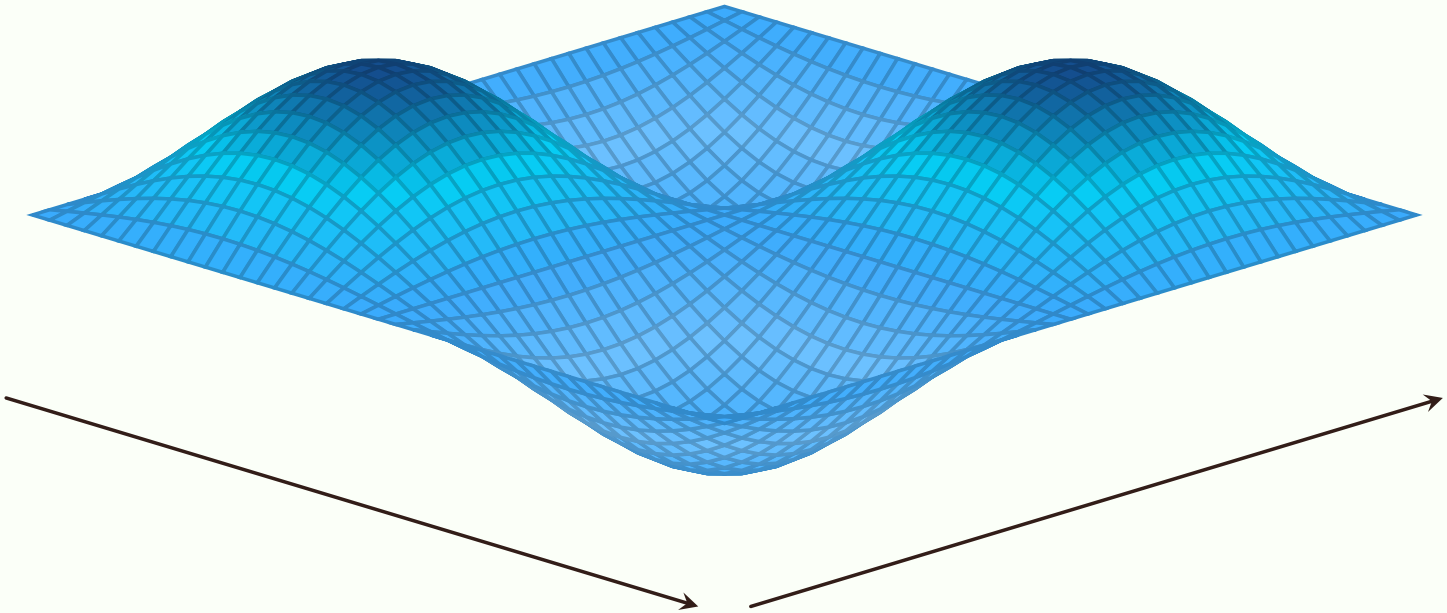
What will I present today?

- Motivation
- Simplicial Complexes
- Persistent Homology
- Current Projects
- Future Work (joint with Siemens on real data)

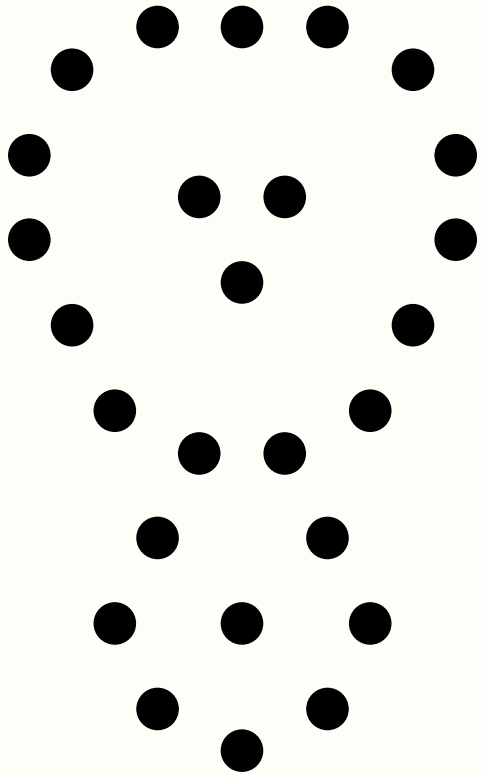
Part I: Motivation

Where does data come from?

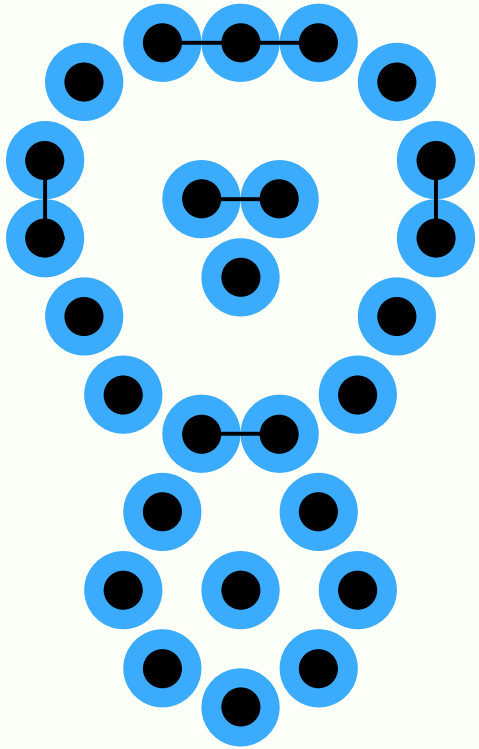
We are interested in the shapes of manifolds.



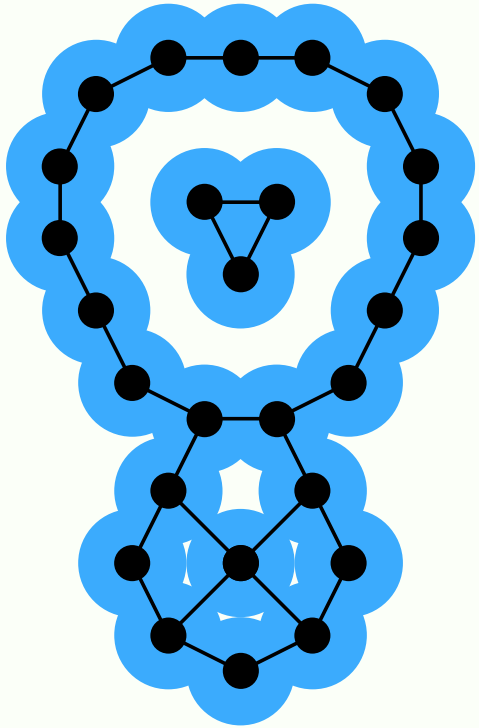
What is the 'shape' of the world where our data comes from?



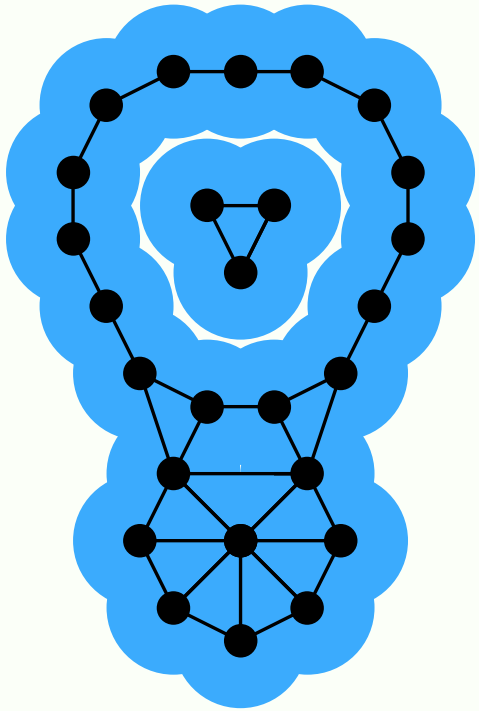
This point cloud has no form in its own right in the technical sense, but we clearly see two ’rings’ connected to each other. In the smaller lower ’ring’ there is a point, in the larger upper ring a ’triangle’.



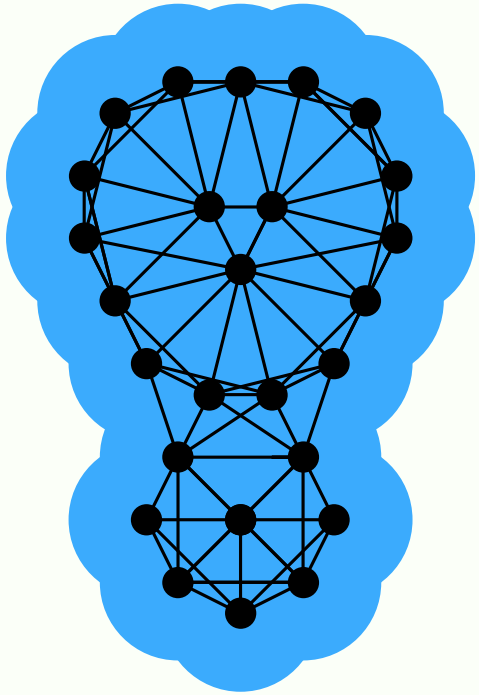
Imagine every point as a balloon. We now inflate this balloon by one $r = 0.2$ and more distinct structures become visible in space.



As soon as two balloons touch, we create an edge between the two center points. This $r = 0.4$.



$r = 0.6$.



This is the 1-skeleton with an $r = 0.8$.

What did we just see?

The formation of a 1-dimensional graph by successively increasing the radius of the balls has revealed the underlying structure.

How can this be formulated more precisely?

Algebraic Topology

This is the area of mathematics which deals with invariant properties of high dimensional geometric objects or spaces.

Simple invariants:

- Dimension: $\mathbb{R}^2 \neq \mathbb{R}^3$ because $2 \neq 3$.
- Determinant: is $\det(A) = \det(B)$ for two matrices A, B, then they are considered to be similar.

We believe that different data comes from different spaces. How can we detect this and sort the data accordingly?

We don’t want to distinguish data only by their ’holes’, therefore we would like to introduce a ’magnifying glass’ which shall allow us to capture the **structure of a space** in different granularity.

Which structures do we observe?

Part II: Simplicial Structures

Real-world multivariate data

Time series, non-analytic or anisotropic data, noise and more ...

- Often unstructured point clouds.
- n items with d attributes gives $n \times d$ matrices.
- Non-random often anisotropic sampling from $\mathbb{R}^d$.

Manifold assumption.

There is a n-dimensional Riemannian manifold M with $n \ll d$ where our data comes from.

Simplex

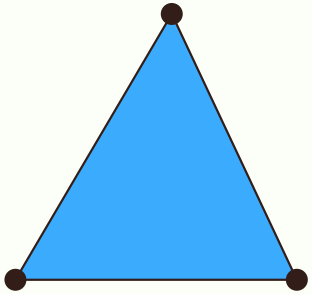
Our structure to work with.

Given a set $X = \{x_0, \ldots, x_k\} \subset \mathbb{R}^d$ of $k+1$ affinely independent points, the k-dimensional simplex Δ spanned by X is the set of convex combinations

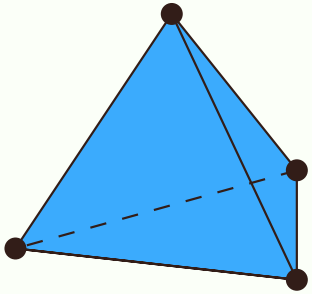
$$\sum_{i=0}^{k} \lambda_i x_i, \quad \text{with} \quad \sum_{i=0}^{k} \lambda_i = 1 \quad \text{and} \quad \lambda_i \geq 0. \tag{1}$$

Δ^0 or 0-simplex

Δ^1 or 1-simplex



Δ^2 or 2-simplex



Δ^3 or 3-simplex

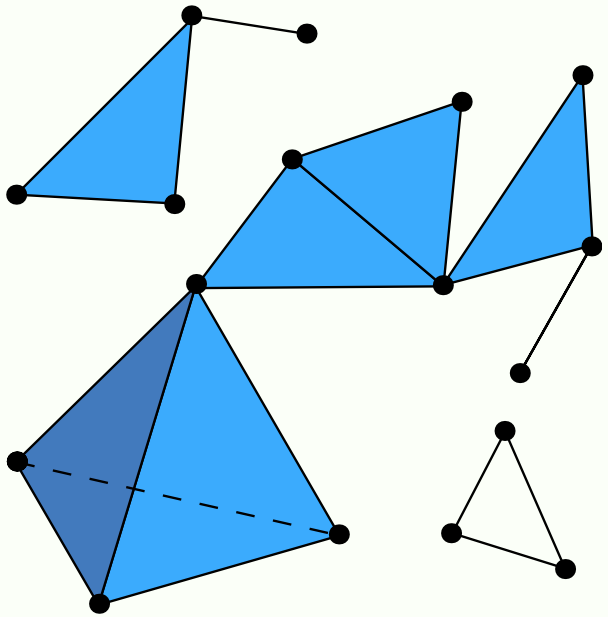
Simplicial complex

Gluing together pieces of increasing dimension.

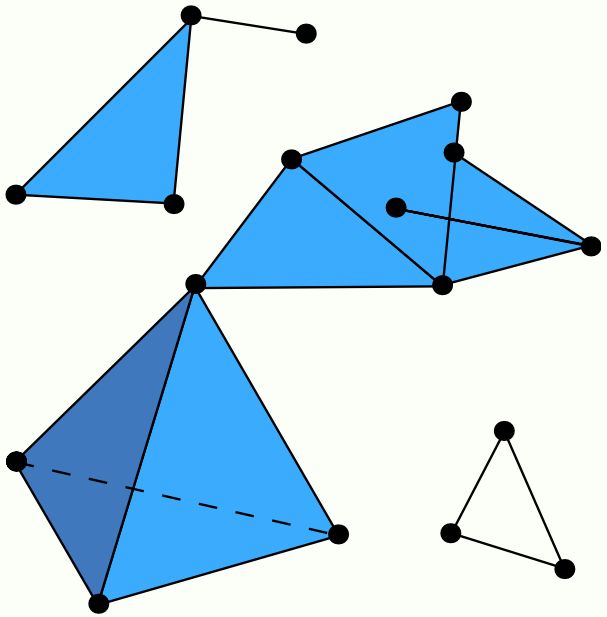
- Gluing together pieces of increasing dimensions.
- Preserve the dimension of the simplex.
- The dimension of the simplicial complex is the dimension of the n-simplex.

A *simplicial complex* K is a topological space realized as a union of any collection of simplices Δ with the following properties:

- Any face of a simplex Δ is also in Δ.
- The intersection of any two simplices of Δ is also in Δ.



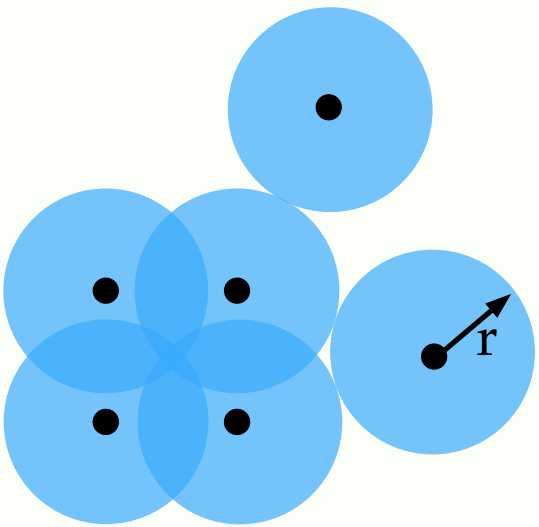
A valid simplicial complex.

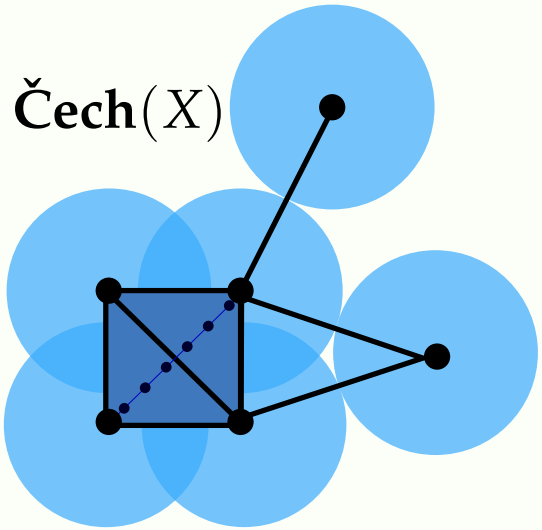


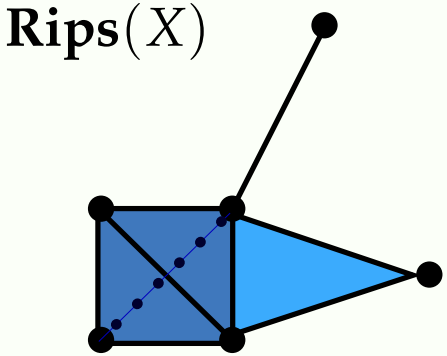
Invalid construction.

Converting data to simplicial complexes

I.e. Čech-, Vietoris-Rips- or witness-complex.







Creating a Čech-complex: Use any distance measure $dist(\cdot, \cdot)$ on your manifold and determine a threshold r. Create an edge between two points if $\text{dist}(p_i, p_j) \leq r$.

- Boundary is easy to describe.
- Geometrically obvious.

The boundary of a line segment are the two endpoints.
The boundary of a triangle is the union of its edges.
The boundary of a tetrahedron is the union of triangular faces.

⋮

Finding Patterns

Boundaries.

- A boundary itself has no boundary.
- Beeing boundariless in finite dimensions coincides with the intuition of a loop.
- We define these things in algebraic terms.

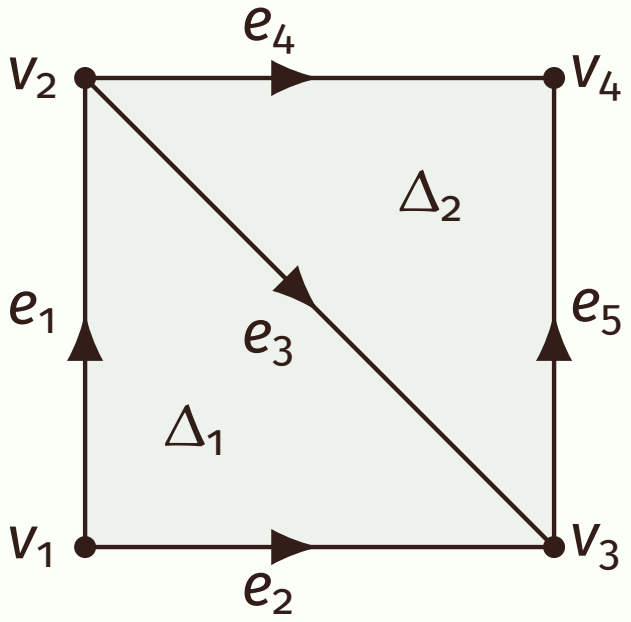
Part III: Calculating $H_k(X)$

Realize Simplices as Algebraic Objects

X_k is the set of k-simplices of the simplicial complex X.
The *chain group* $C_k(X)$ is the $\mathbb{Q}$-vector space with X_k as a basis.

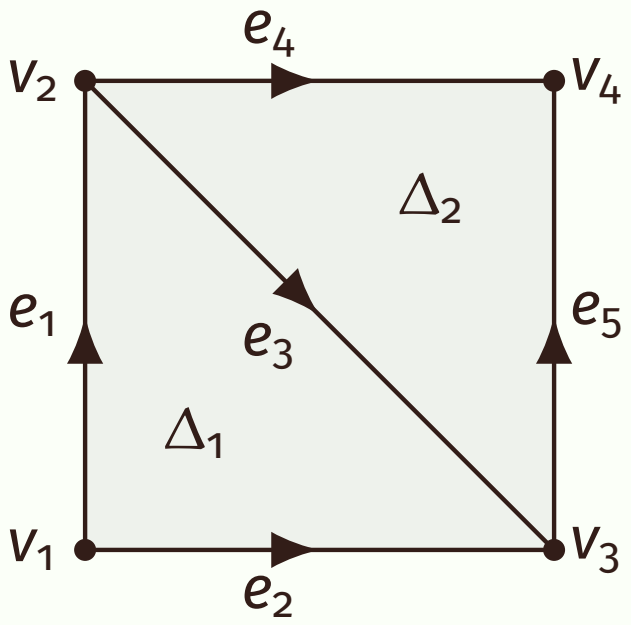
Chain groups

Let Δ^k and $\Delta^{k\prime}$ be two k-simplices. Then we define a bunch of new 'chains' as all possible sums and scalar multiples of the simplices, i.e. $\lambda_1\Delta^k + \lambda_2\Delta^{k\prime}$. This is denoted in general as $\text{span}(\Delta_1^k, \ldots, \Delta_n^k)$.



$C_0(X)$ is the span of the vertices $\text{span}(v_1, v_2, v_3, v_4)$. Having two copies of v_2 and four of v_3 yields this algebraic equation $2v_2 + 4v_3$.

There are meaningless equations such as $\frac{1}{2}v_4 - v_2 - \frac{3}{4}v_5$ which exist for their own right.



$C_1(X)$ is therefore the span of the set $\{e_1, e_2, e_3, e_4, e_5\}$. We can talk about a path, as the edges are line segments.

Again some realizations are geometrically ridiculous: $-e_5 + e_4 + e_1 + e_3$.

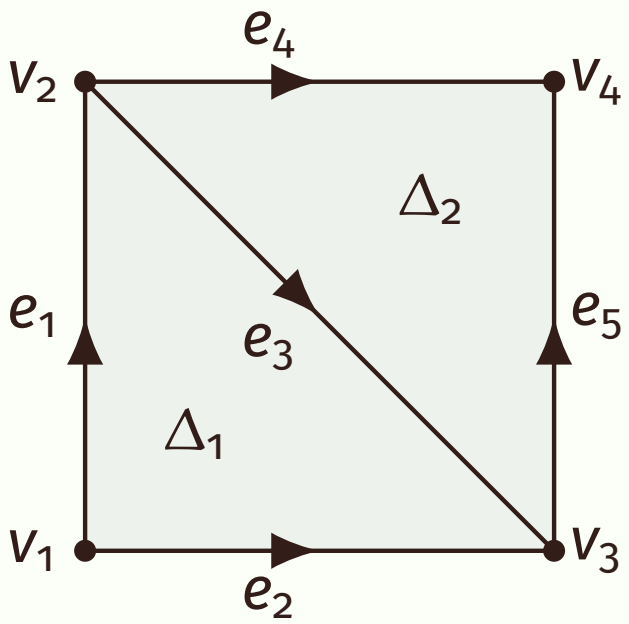
Extending to Higher Dimensions

- The chain group $C_2(X) = \text{span}(\Delta_1, \Delta_2)$.
- What is a path of triangles?
- Geometric analogies become tenuous in high dimensions.
- Algebra helps. Find low dimensional intuitions and generalize them.

The Boundary Operator

- Chain groups are linear groups.
- Their elements are vector spaces.
- They induce linear maps $\partial : C_k(X) \to C_{k-1}(X)$.
- Single vertices are boundariless, i.e. $\partial(v) = 0$.

The Boundary Operator



$$\begin{aligned}
\partial(v_1, v_2) &= v_2 - v_1 \\
\partial(e_1 + e_4 - e_5 - e_3) &= \partial e_1 + \partial e_4 - \partial e_5 - \partial e_3 = \\
&= (v_2 - v_1) + (v_4 - v_2) - (v_4 - v_3) - (v_3 - v_2) = \\
&= v_2 - v_1.
\end{aligned}$$

The Boundary Operator

Results.

- The endpoint of v_2 minus the starting point of v_1. Each successive edge cancels out the ending vertex of the edge before it.
- Each starting edge cancels the starting edge after it.
- In case of a loop the boundary operator becomes 0.
- The boundary operator is given by an alternating sum.

The *boundary operator* or sometimes *boundary map* $\partial_k : C_k(X) \to C_{k-1}(X)$ is given by:

$$\partial_k(\Delta^k) = \sum_{i=0}^{k} (-1)^i \Delta^k \Big|_{[v_0,\ldots,v_{i-1},\hat{v}_i,v_{i+1},\ldots,v_k]} \tag{2}$$

Observations

- Boundary operators can be composed: $\partial \circ \partial = \partial^2$.
- Every chain which is a boundary of higher-dimensional chains is boundariless, i.e. $\partial^{2\mathbb{Z}} = 0$.
- Low dimensional examples make sense and can be extended.
- Boundariless 'cycles' correspond to loops.

Observations

... in the language of algebraic topology.

- The kernel is defined as
  $\ker \partial_k := \partial_k^{-1}(e_{C_{k-1}(X)}) = \{z \in C_k(X) : \partial_k(z) = e_{C_{k-1}(X)}\}$.
- A k-cycle is then defined as $Z_k(X) = \ker(\partial_k) \subset C_k(X)$.
- The k-boundaries $B_k(X) = \mathrm{im}(\partial_{k+1}) \subset Z_k(X)$.

$$\cdots \xrightarrow{\partial_{k+1}} C_k \xrightarrow{\partial_k} C_{k-1} \xrightarrow{\partial_{k-1}} \cdots \xrightarrow{\partial_2} C_1 \xrightarrow{\partial_1} C_0 \xrightarrow{\partial_0} 0 \xrightarrow{\partial_{k+1}} \cdots \quad (3)$$

Simplicial Homology Group

The k^{th} homology group H_k is a quotient group, defined by 'removing' cycles that are boundaries from a higher dimension:

$$H_k = Z_k / B_k = \ker \partial_k / \operatorname{im} \partial_{k+1} \tag{4}$$

This definition gives rise to the k^{th} Betti-number:

$$\beta_k = \operatorname{rank} H_k \tag{5}$$

Intuitively: Calculate all boundaries, remove the boundaries that come from higher-dimensional objects, and count what is left.

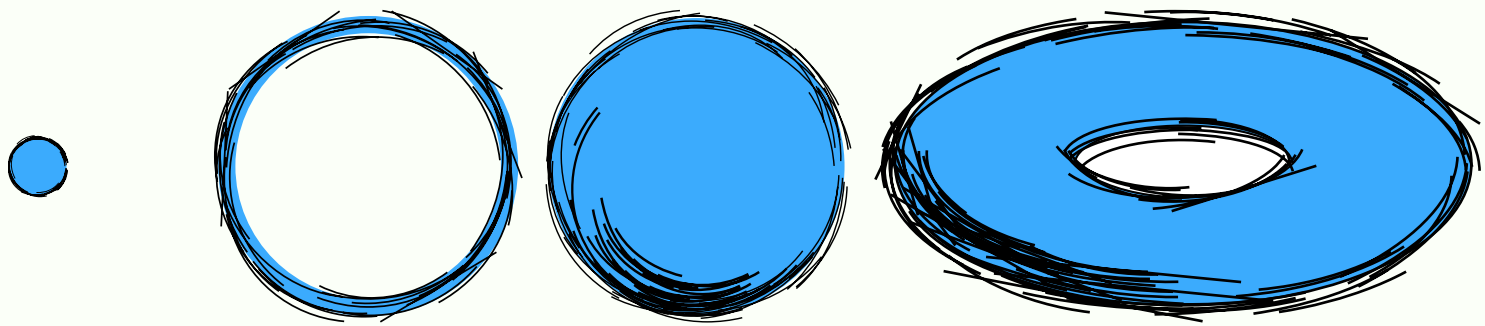
Betti numbers

Another topological invariant.

Betti numbers 'count' the number of holes in the dimensions that occur in space.

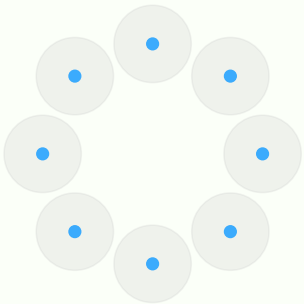
β_0	Connected components
β_1	Tunnels
β_2	Voids
⋮	⋮

Space	β_0	β_1	β_2
Point	1	0	0
Circle	1	1	0
Sphere	1	0	1
Torus	1	2	1

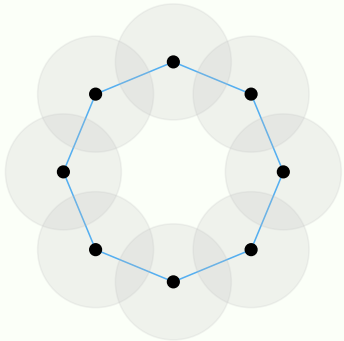


Betti numbers

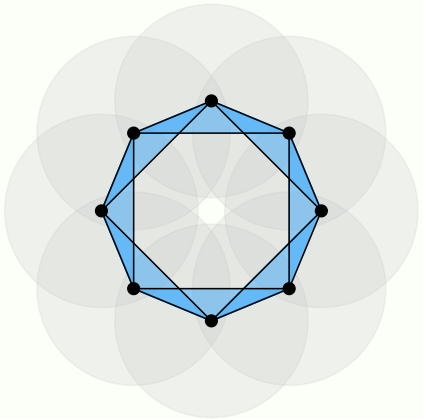
Another topological invariant.



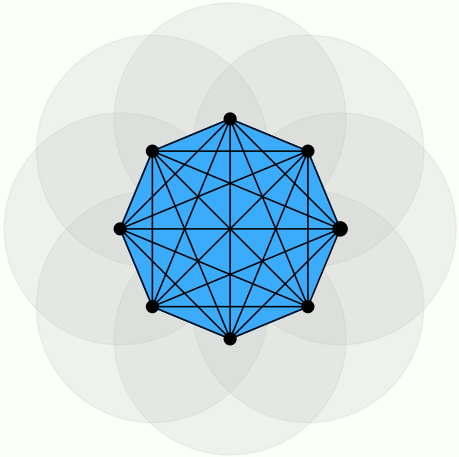
(a) $r = 0.2$.



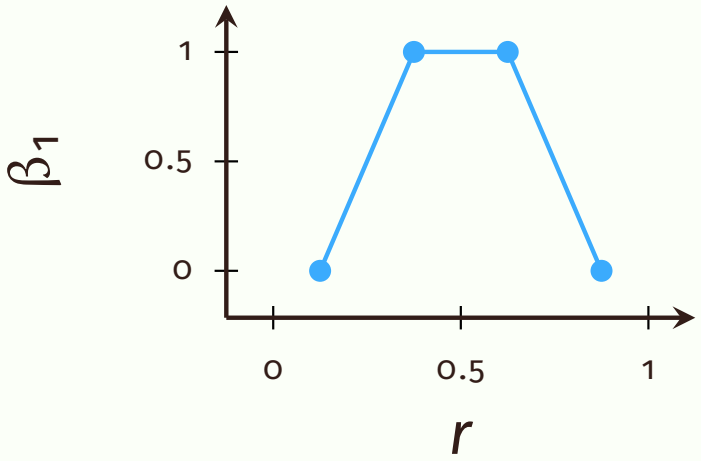
(b) $r = 0.4$.



(c) $r = 0.6$.

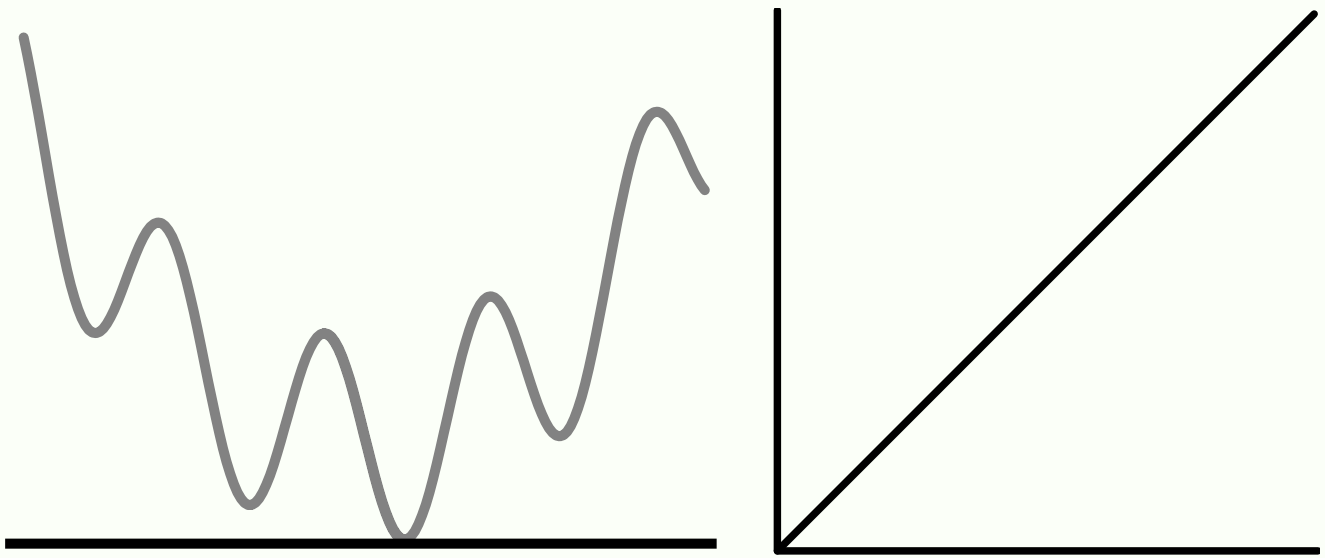


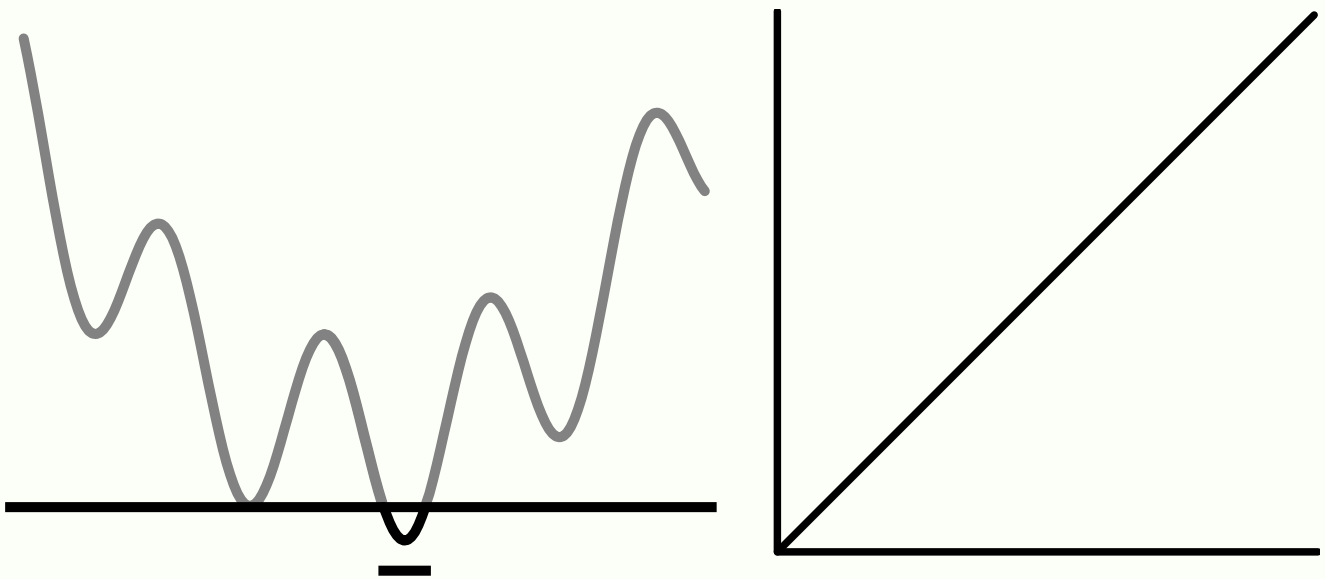
(d) $r = 0.8$.

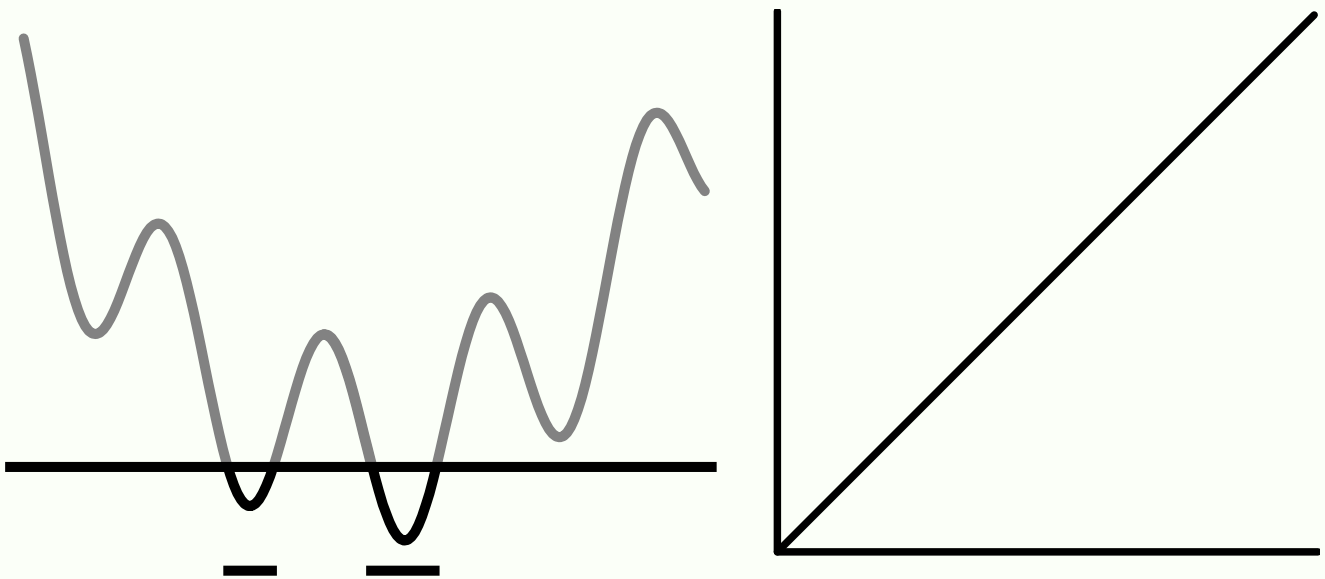


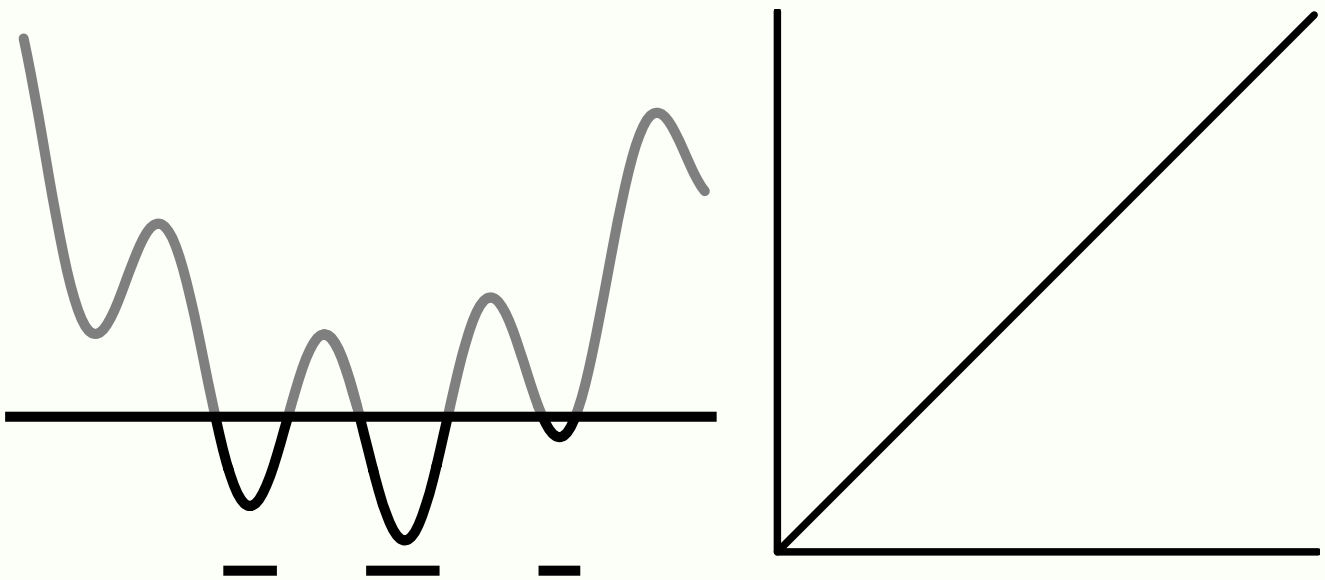
Persistent Homology

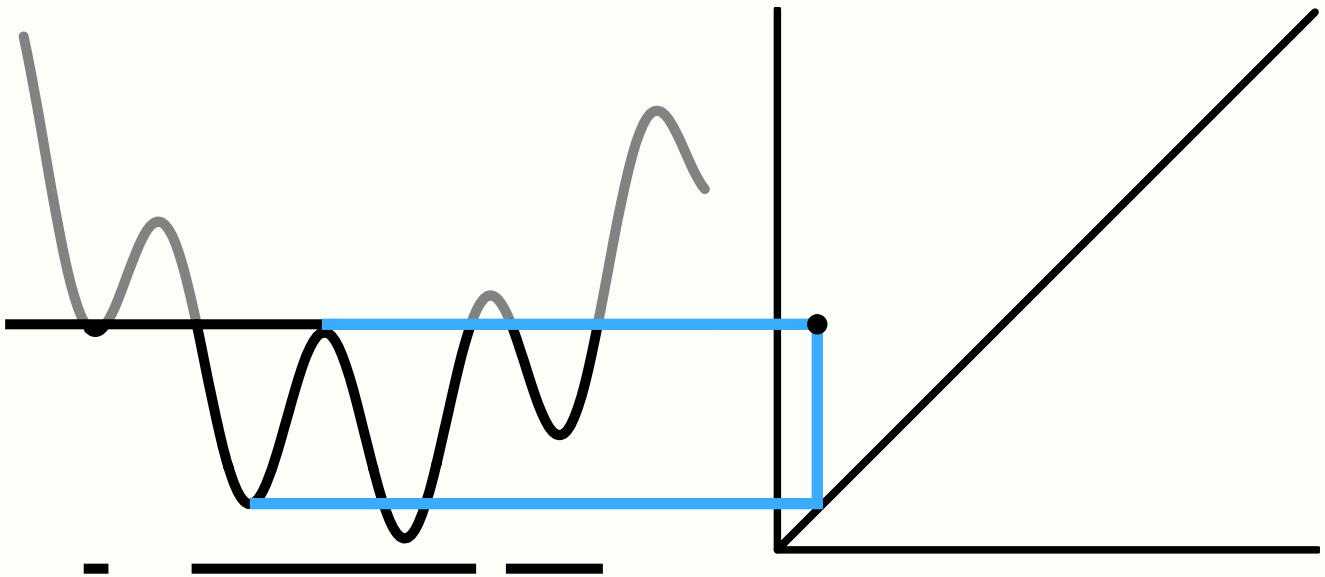
- Calculates the k-th Betti numbers.
- Gives an overview of the evolving topology of a filtration.
- Robust against noise.
- Detects dense regions in data.

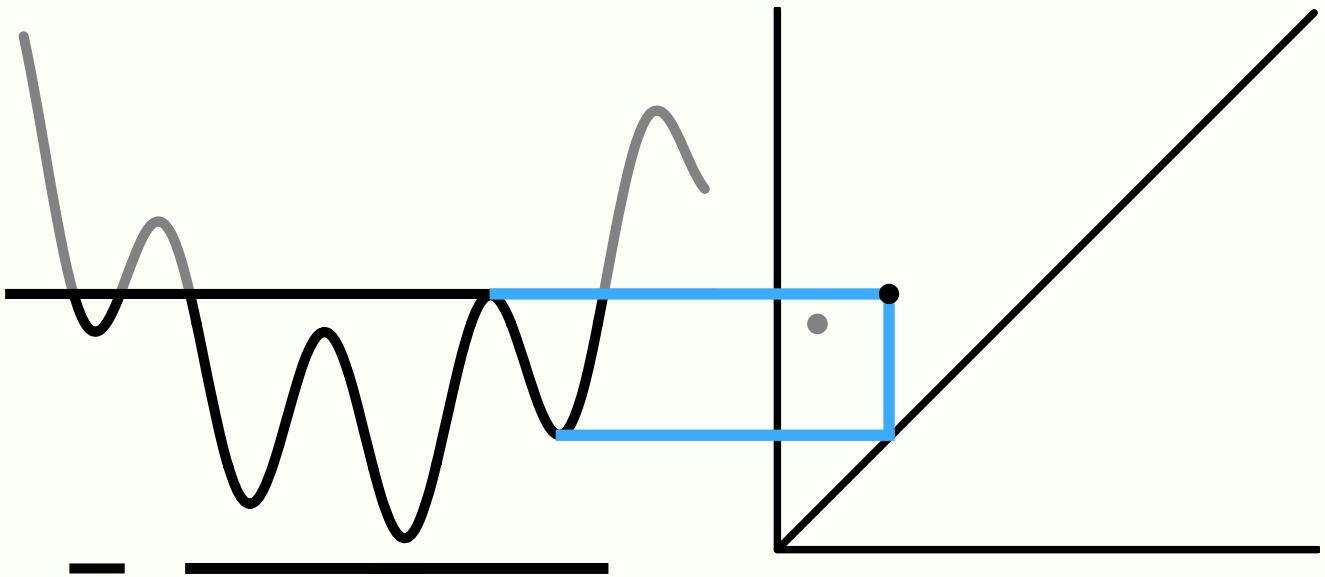


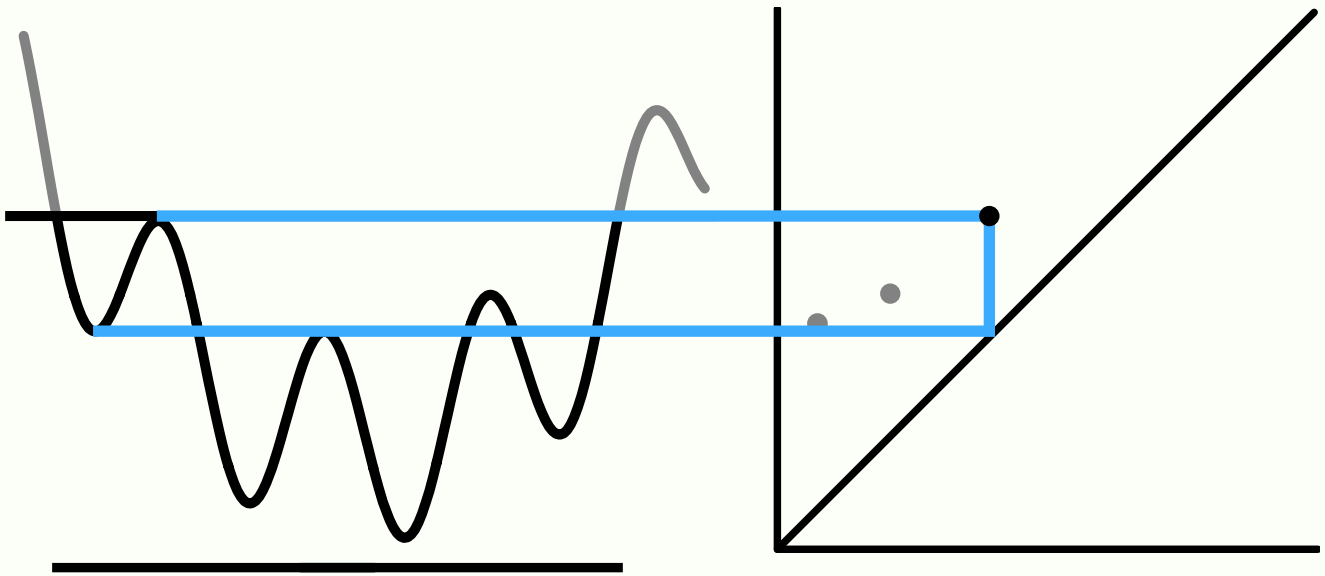


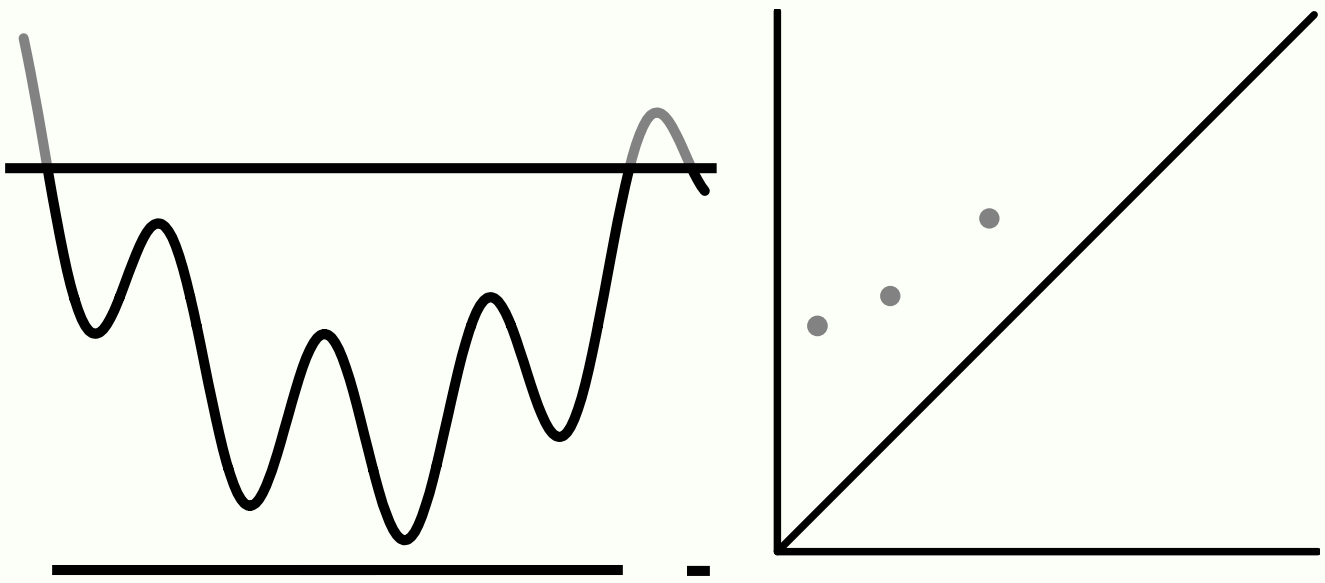


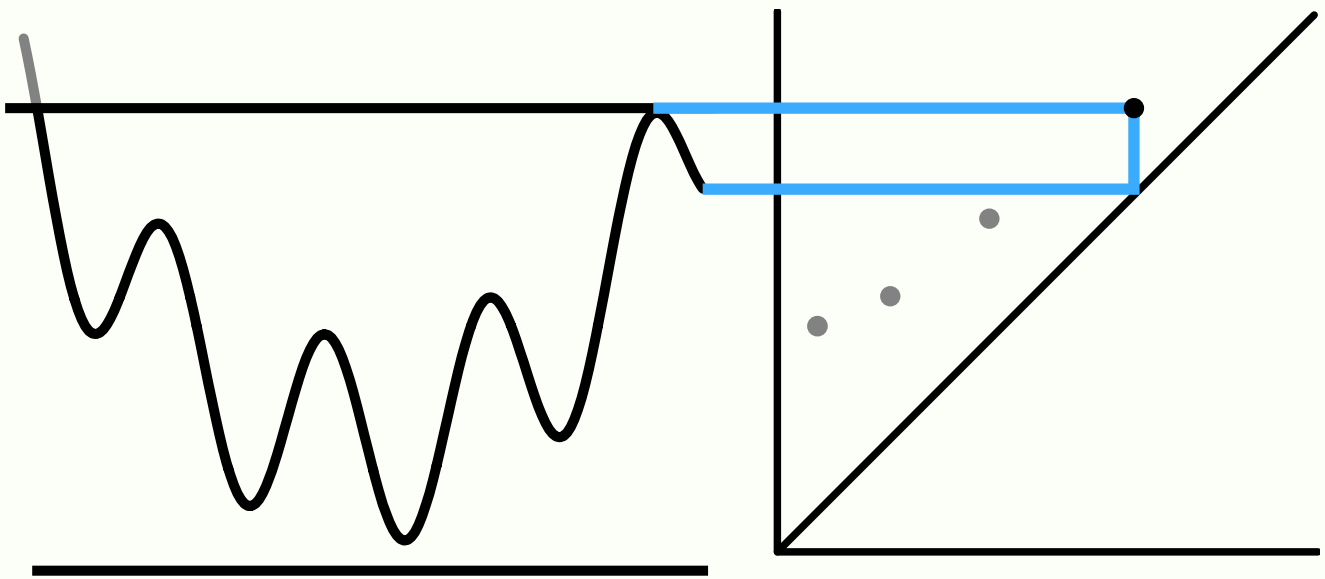


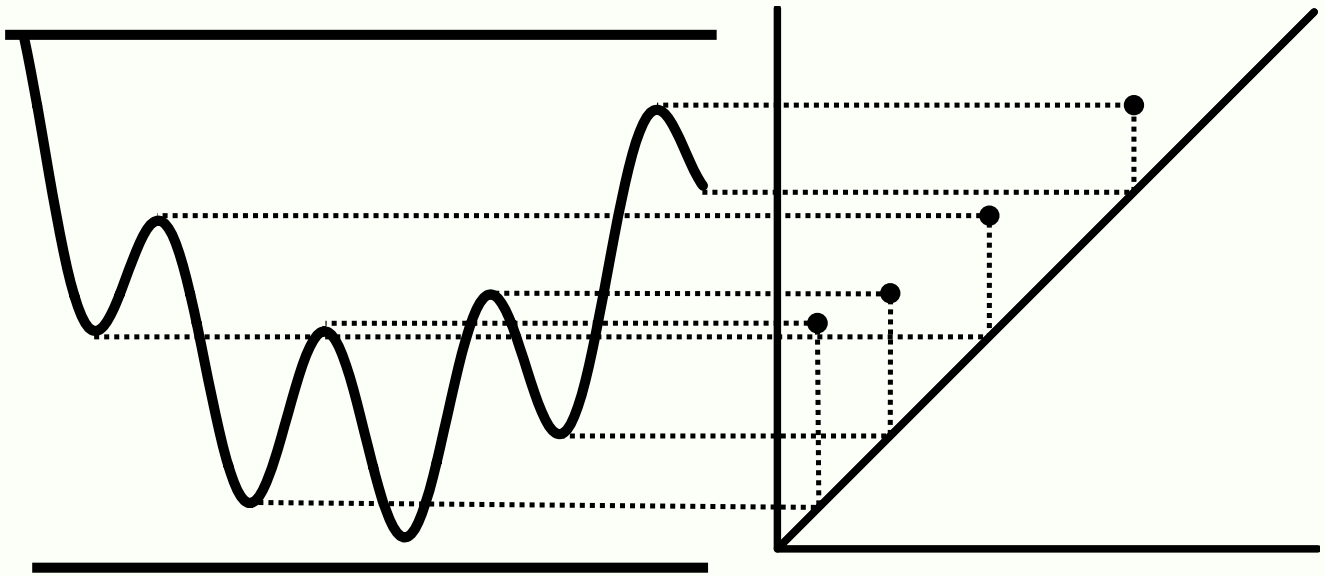


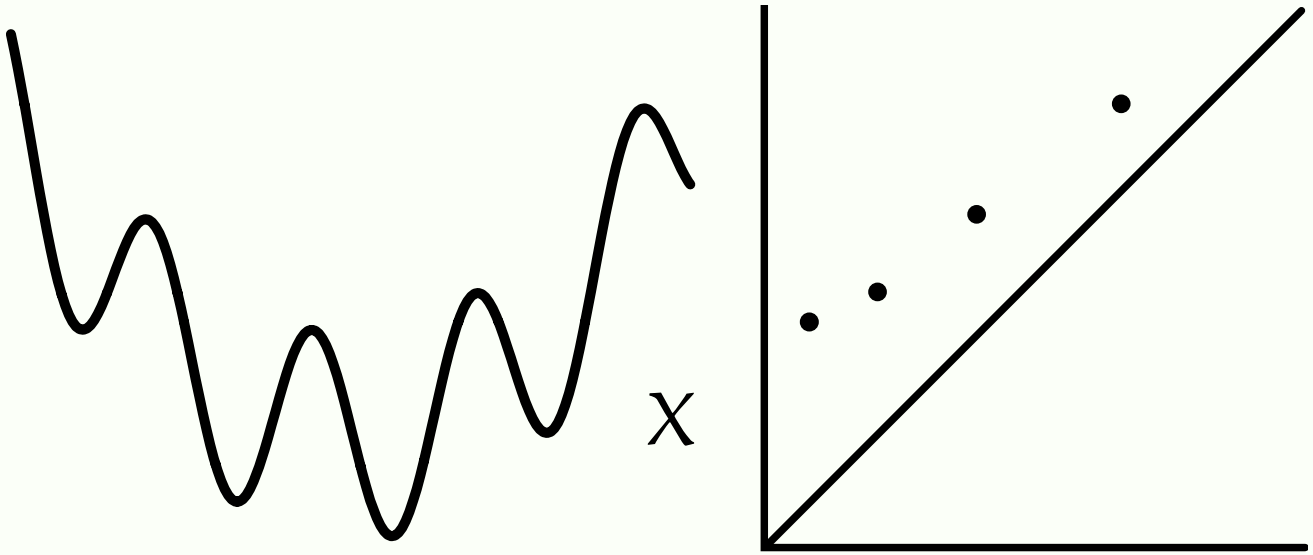








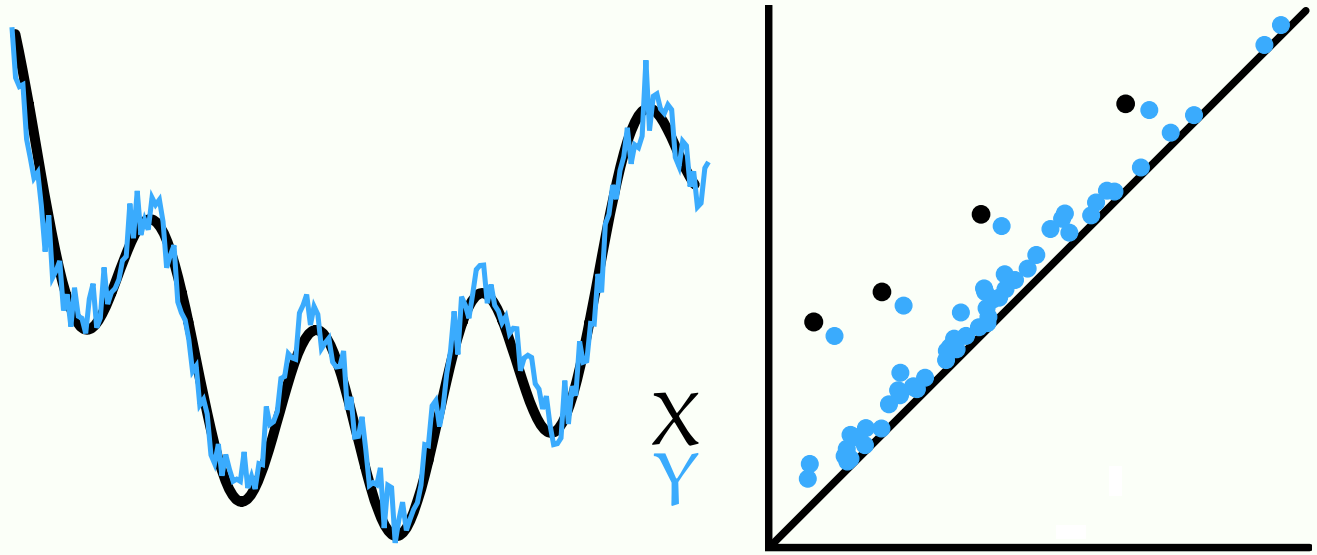




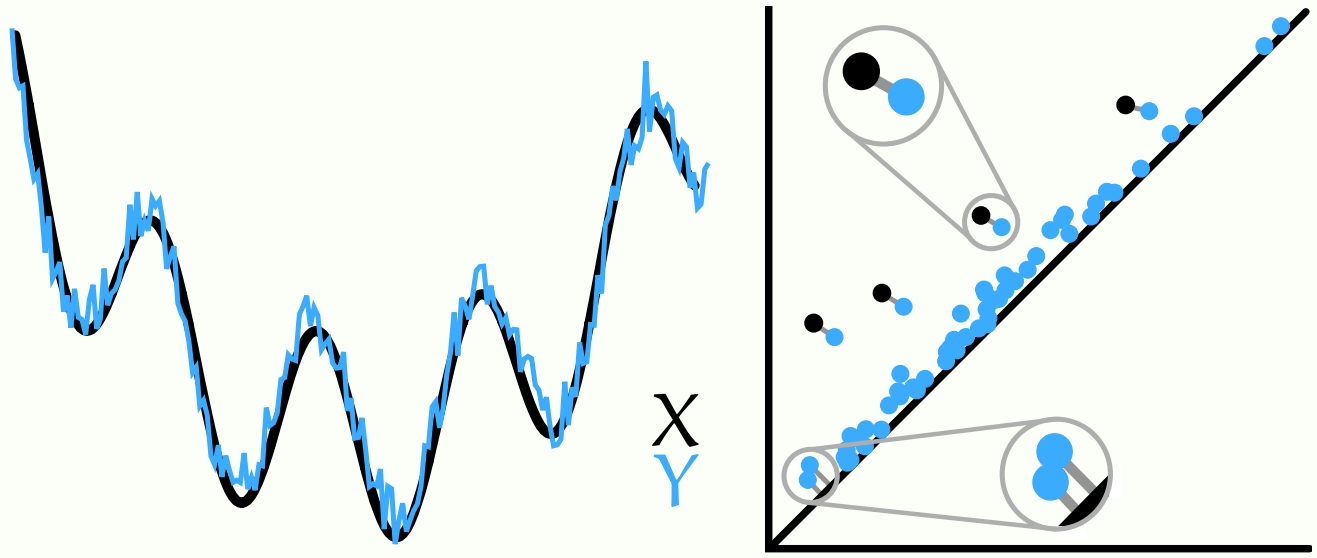
A persistence diagram of a curve.

Persistent Homology

Persistence diagram by Rieck.



Persistent homology is robust against noise in a very precise way.

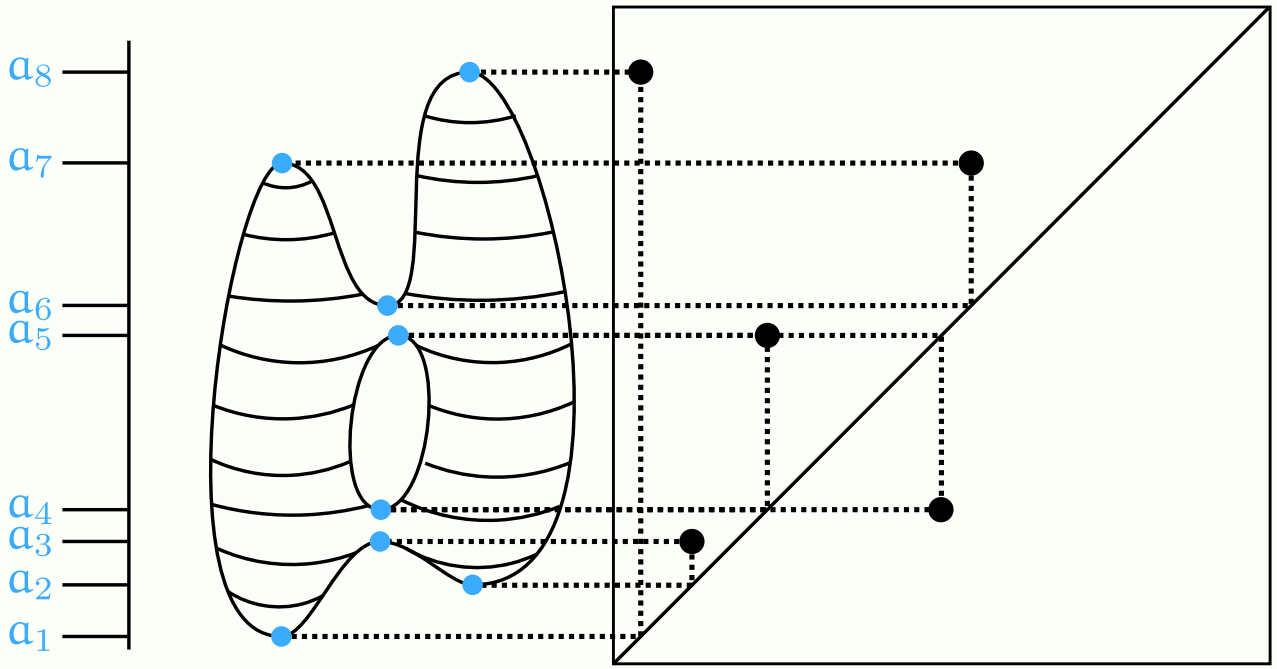


The **Bottleneck** distance compares two persistence diagrams:

$$d_B(\mathcal{P}_1, \mathcal{P}_2) = \inf_\varphi \sup_{x \in \mathcal{P}_1} \|x - \varphi(x)\|_\infty$$

Morse Theory

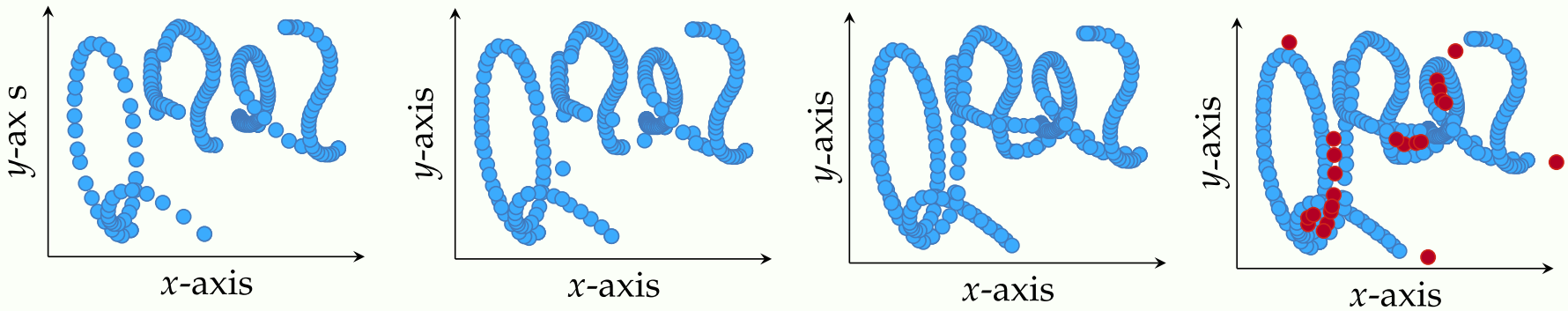
Connection to differential topology.

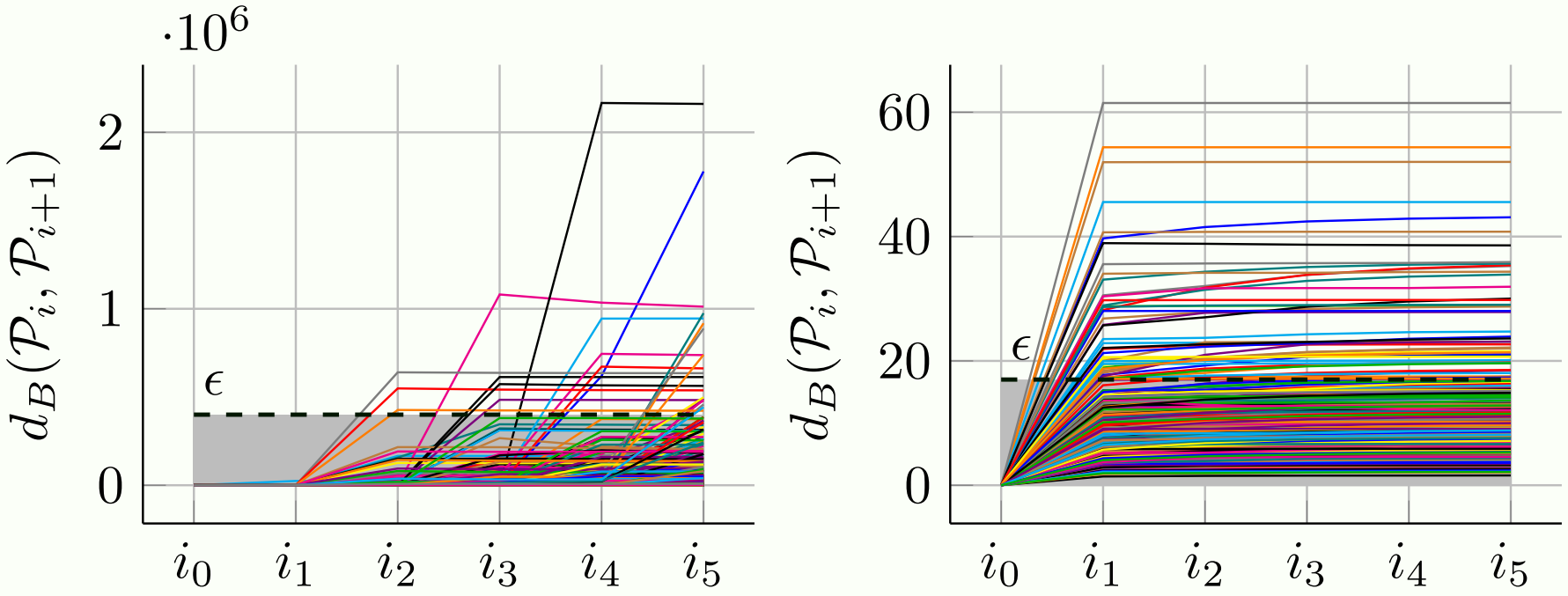


Minima in hyperplanes create new 'connected' components.
Maxima in turn destroy them by merging connected components.
Saddle points either create cavities or 'holes' or connect two connected components.

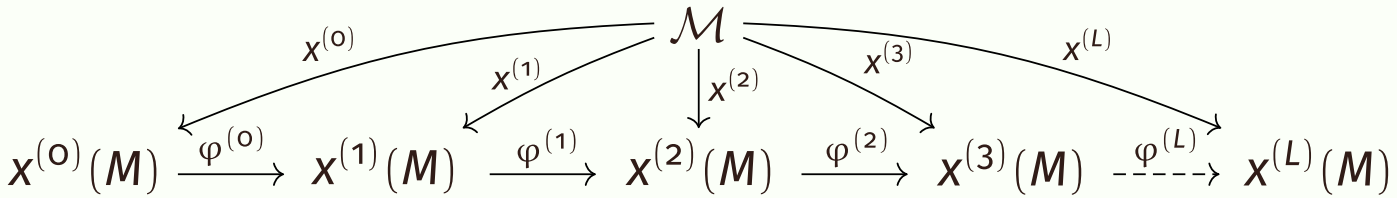
Part V: Current Projects

- Method of Natural Neighbors for interpolation.
- Topological plausible interpolation.
- Stabilizing the method with topological stopping.
- Interpolation performs until topology remains 'the same'.





Natural Neighbor interpolation with heuristic stop ϵ.
Different filtrations lead to impressively similar results, but are dependend on parametrization.



- **Manifold assumption**:
  Data lives on low dimensional manifold.
- What is the dimension of this manifold $\mathcal{M}$?
- How 'complex' is the topology of the manifold $\mathcal{M}$?
- How many dimension do I really need for trajectory $\varphi^{(l)}$?

Current Projects

Signature Identification joint with Michael Nissen.

- Visualization of the topology of signatures.
- Knowledge inference from persistent homology.
- Choosing appropriate filtrations.
- Interpolation of the data.
- Parametrizing neural networks with knowledge from persistent homology.
- Using residual networks to identify signatures.

Part IV: Future Work

Motivation

Beyond mathematical logic.

Applying algebraic topology and category theory yields the category of databases as theoretical framework. Four good reasons for category theory:

- Morphisms preverse the database structure.
- Categorical constructions (limits or colimits) result in database operations (union or join).
- Restructuring and integrating data becomes easy.
- The schema has a natural geometrical structure.

Following the work of **David Spivak** a relation can be realized as simplicial complex. Now the following topics can be posed, which I'll try to elaborate:

- Application of Spivaks construction to real world data.
- Visualization of functional dependencies using TDA.
- Triangulation of relations capturing FDs.
- Clustering of large datasets according to this triangulation.

References

Jean-Daniel Boissonnat, Frédérik Chazal, Mariette Yvinec
Geometric and Topological Inference.
Cambridge, 2018.

David Spivak
Simplicial Databases.
Arxiv abs:0904.2012, 1–25, 2009.

Bastian Rieck
Persistent homology in multivariate data visualization.
University of Heidelberg, pages 1–307, 2017.

Bastian Rieck
A Primer in Persistent Homology.
URL: https://www.math.uni-hamburg.de/.

Jeremy Kun
Homology Theory - A Primer.
URL: https://jeremykun.com/.

Acknowledgement

I would like to thank **Justin Noel** and **Jan Frahm** for bringing me closer to algebraic topology and differential geometry.

Special thanks to **Bastian Rieck**, who provided me some of his graphics for this presentation with friendly permission.

Furthermore I would also like to thank my supervisor **Richard Lenz**, who breaks up complex contexts and makes me bring them into an orderly form.