A primer in persistent homology

Bastian Rieck

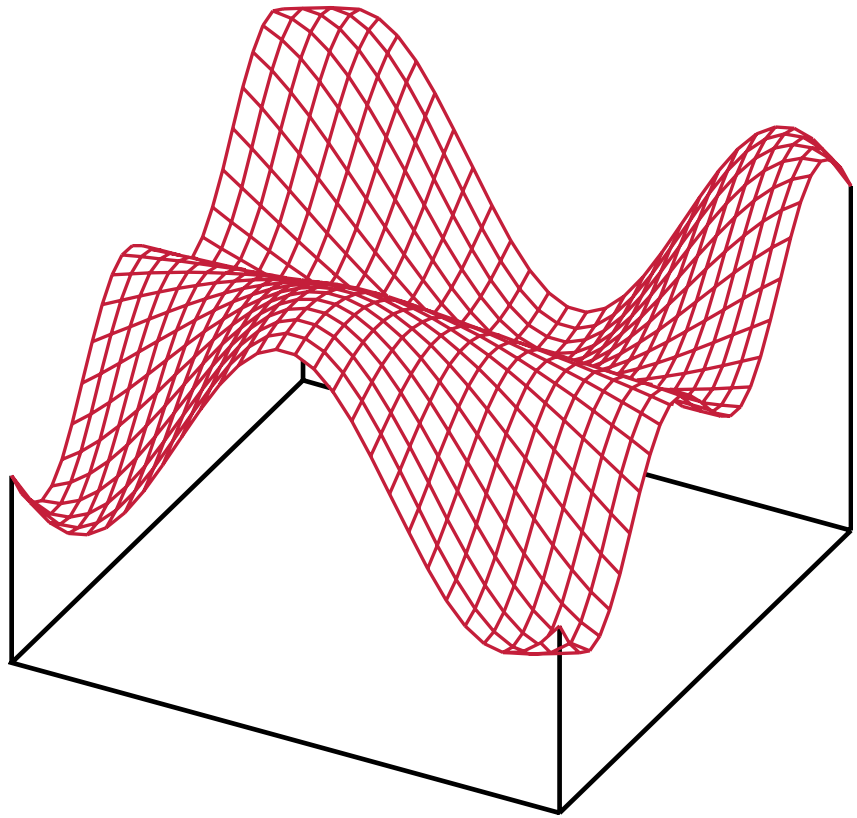
UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386

HGS
MathComp

TECHNISCHE UNIVERSITÄT
KAISERSLAUTERN

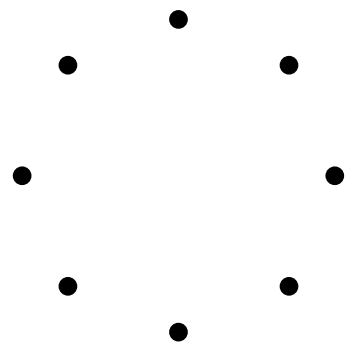
Motivation

What is the ‘shape’ of data?



A simple example

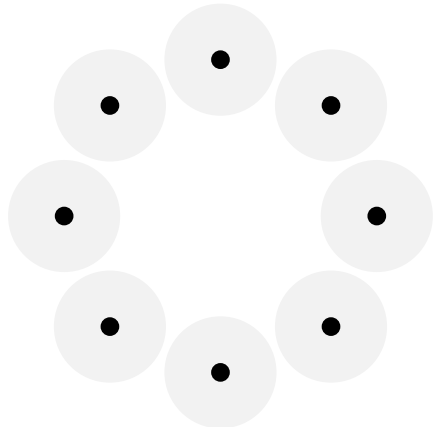
What is the shape of this set of points?



Technically, a set of points does not have a ‘shape’. Still, we *perceive* the points to be arranged in a circle. How can we quantify this?

A simple example

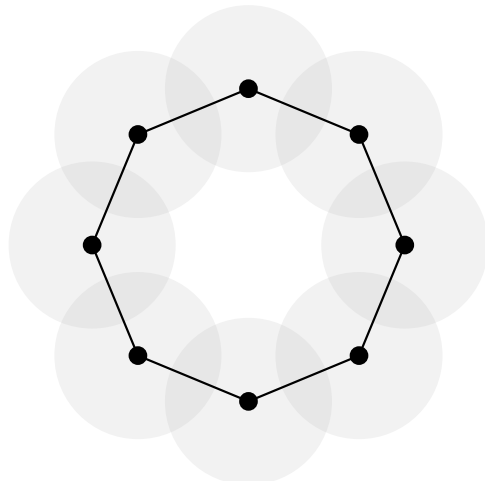
What is the shape of this set of points?



We can ‘squint’ our eyes and look at how the connectivity of the points changes. The more we squint, the more connections we see.

A simple example

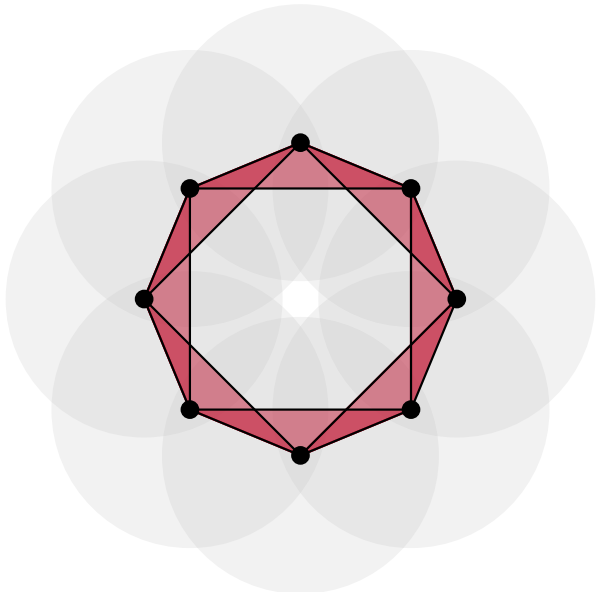
What is the shape of this set of points?



We can ‘squint’ our eyes and look at how the connectivity of the points changes. The more we squint, the more connections we see.

A simple example

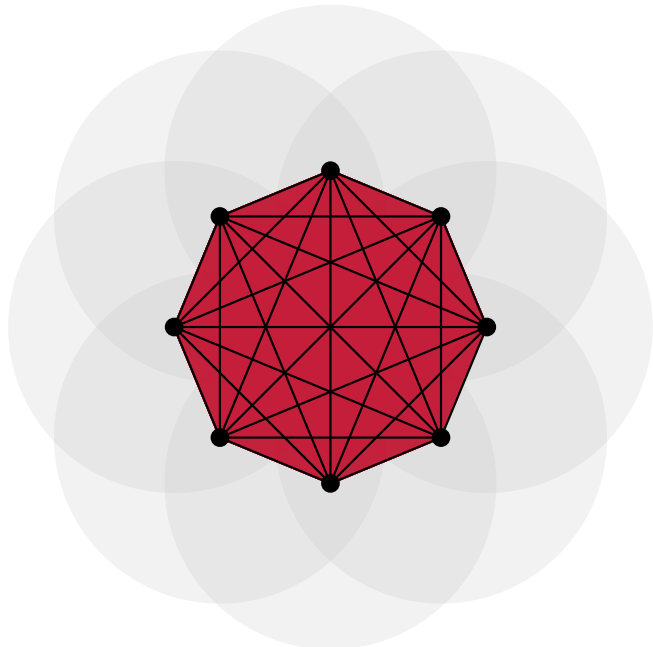
What is the shape of this set of points?



We can ‘squint’ our eyes and look at how the connectivity of the points changes. The more we squint, the more connections we see.

A simple example

What is the shape of this set of points?



We can ‘squint’ our eyes and look at how the connectivity of the points changes. The more we squint, the more connections we see.

What did we see?

Points are arranged in a circle, as long as the radius of the disks we use to cover them does not exceed a certain critical threshold.

How can we formulate this more precisely?

Algebraic topology

The branch of mathematics that is concerned with finding *invariant properties* of high-dimensional objects.

Simple invariants

1. Dimension: $\mathbb{R}^2 \neq \mathbb{R}^3$ because $2 \neq 3$
2. Determinant: If matrices A and B are similar, their determinants are equal.

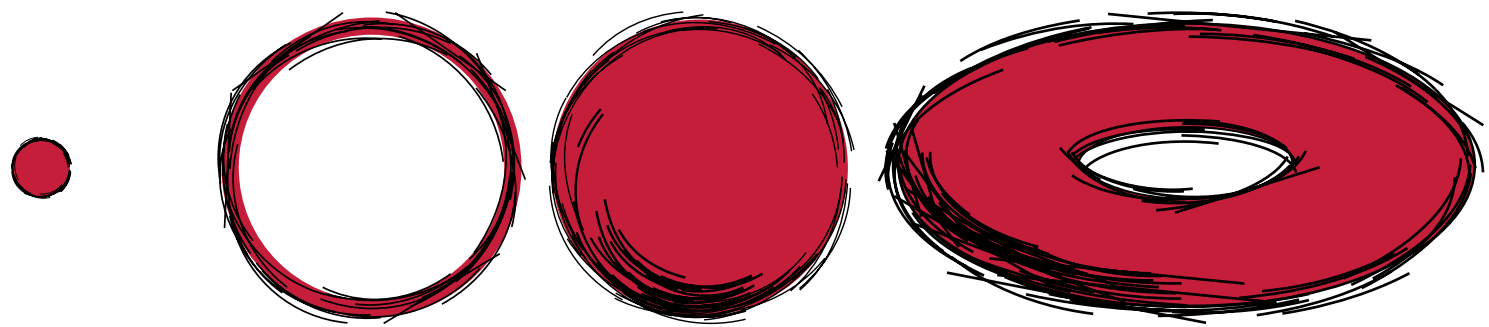
Betti numbers

A topological invariant

Informally, they count the number of holes in different dimensions that occur in an object.

β_0	Connected components
β_1	Tunnels
β_2	Voids
$\vdots$	$\vdots$

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



Calculating Betti numbers

The k^{th} Betti number β_k is the rank of the k^{th} *homology group* $H_k(X)$ of the topological space X.

To define this formally, we require a notion of 'holes' in simplicial complexes. This, in turn, requires the concepts of boundaries and cycles.

Technically, I should write *simplicial homology group* every time. I am not going to do this. Instead, let us first talk about *simplicial complexes*.

Simplicial complexes

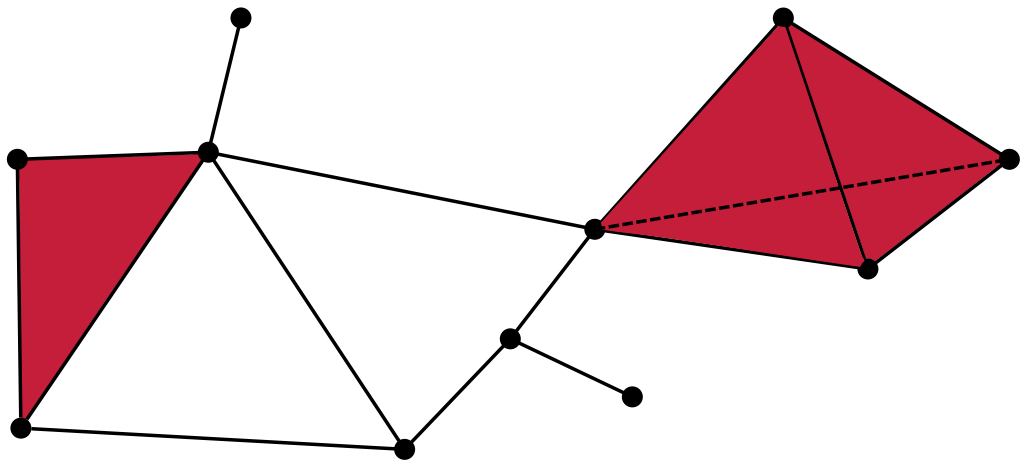
A family of sets K with a collection of subsets S is called an *abstract simplicial complex* if:

1. $\{v\} \in S$ for all $v \in \mathrm{K}$.
2. If $\sigma \in S$ and $\tau \subseteq \sigma$, then $\tau \in \mathrm{K}$.

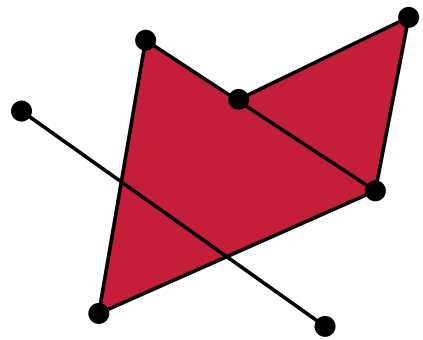
The elements of a simplicial complex are called *simplices*. A k-simplex consists of $k+1$ indices.

Simplicial complexes

Example



Valid



Invalid

Chain groups

Given a simplicial complex K, the p^{th} chain group C_p of K contains all linear combinations of p-simplices in the complex. Coefficients are in $\mathbb{Z}_2$, hence all elements of C_p are of the form $\sum_j \sigma_j$, for $\sigma_j \in \mathrm{K}$. The group operation is addition with $\mathbb{Z}_2$ coefficients.

We need chain groups to algebraically express the concept of a *boundary*.

Boundary homomorphism

Given a simplicial complex K, the p^{th} boundary homomorphism is the homomorphism that assigns each simplex $\sigma = \{v_0, \ldots, v_p\} \in \mathrm{K}$ to its boundary:

$$\partial_p \sigma = \sum_i \{v_0, \ldots, \hat{v}_i, \ldots, v_k\}$$

In the equation above, $\hat{v}_i$ indicates that the set does *not* contain the i^{th} vertex. The function $\partial_p : C_p \to C_{p-1}$ is thus a homomorphism between the chain groups.

Fundamental lemma & chain complex

For all p, we have $\partial_{p-1} \circ \partial_p = 0$: *Boundaries do not have a boundary themselves*. This leads to the *chain complex*:

$$0 \xrightarrow{\partial_{n+1}} C_n \xrightarrow{\partial_n} C_{n-1} \xrightarrow{\partial_{n-1}} \dots \xrightarrow{\partial_2} C_1 \xrightarrow{\partial_1} C_0 \xrightarrow{\partial_0} 0$$

Cycle and boundary groups

$$\text{Cycle group } Z_p = \ker \partial_p$$
$$\text{Boundary group } B_p = \operatorname{im} \partial_{p+1}$$

We have $B_p \subseteq Z_p$ in the group-theoretical sense. In other words, every boundary is also a cycle.

Homology groups & Betti numbers

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

$$H_p = Z_p / B_p = \ker \partial_p / \operatorname{im} \partial_{p+1},$$

With this definition, we may finally calculate the p^{th} Betti number:

$$\beta_p = \operatorname{rank} H_p$$

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

Real-world multivariate data

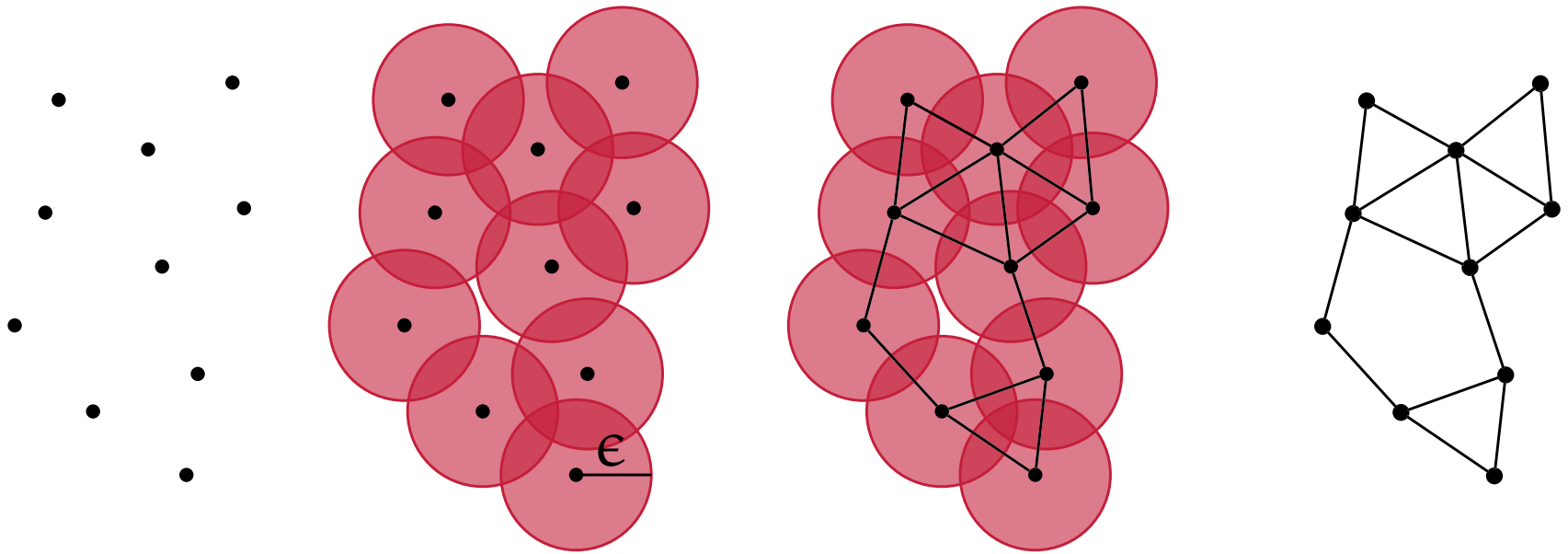
- Often: Unstructured point clouds
- n items with D attributes; $n \times D$ matrix
- Non-random sample from $\mathbb{R}^D$

Manifold hypothesis

There is an unknown d-dimensional manifold $\mathbb{M} \subseteq \mathbb{R}^D$, with $d \ll D$, from which our data have been sampled.

Converting unstructured data into a simplicial complex

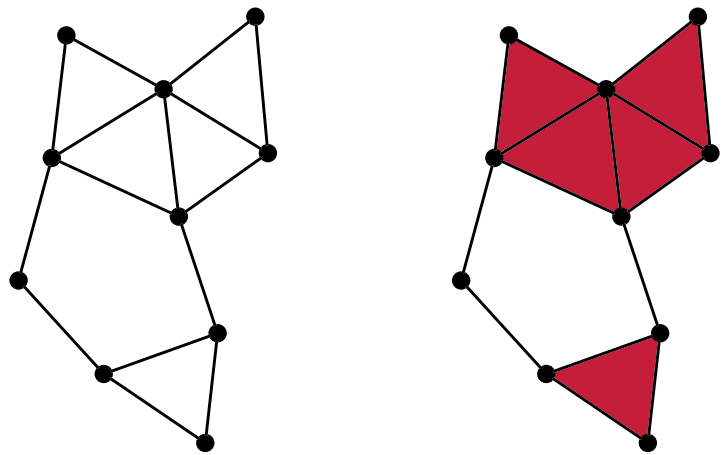
Rips graph $\mathcal{R}_\epsilon$



Use a distance measure $\mathrm{dist}(\cdot,\cdot)$ such as the Euclidean distance and a threshold parameter ϵ. Connect u and v if $\mathrm{dist}(u,v) \leq \epsilon$.

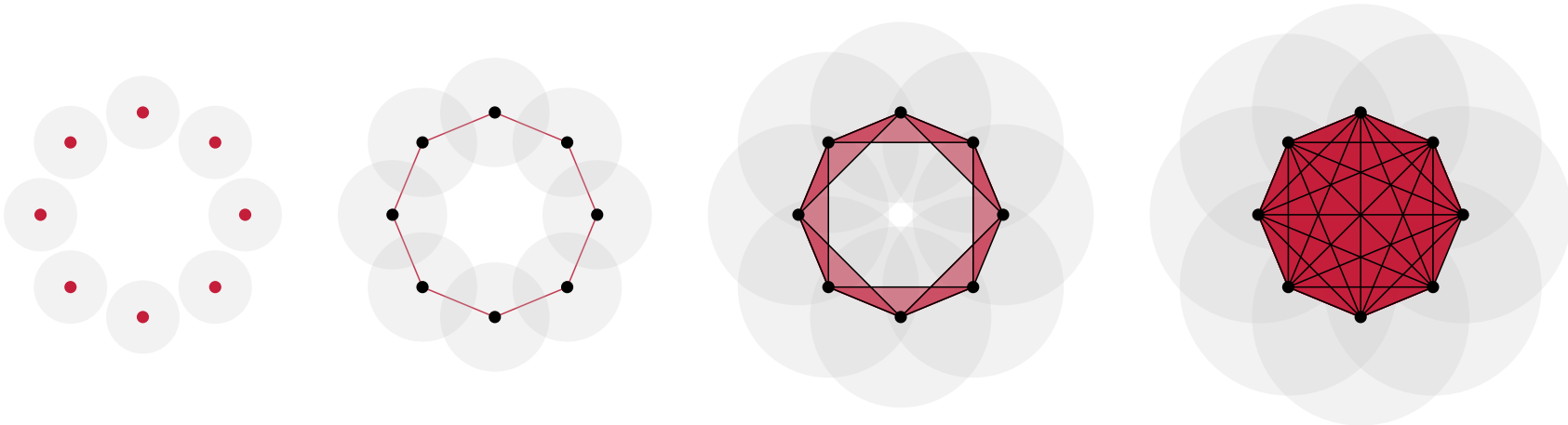
How to get a simplicial complex from $\mathcal{R}_\epsilon$?

Construct the Vietoris–Rips complex $\mathcal{V}_\epsilon$ by adding a k-simplex whenever all of its $(k-1)$-dimensional faces are present.

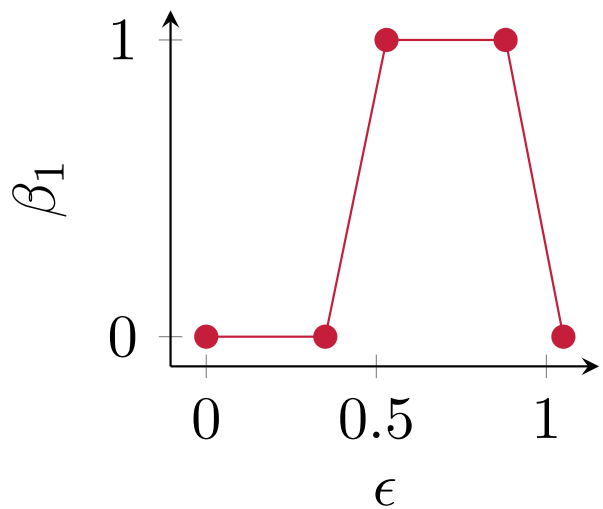


How to calculate Betti numbers?

Direct calculations are *unstable*



$\epsilon = 0.35$ $\epsilon = 0.53$ $\epsilon = 0.88$ $\epsilon = 1.05$



Persistent homology

Note that the 'correct' Betti number of the data *persists* over a certain range of the threshold parameter ϵ. To formalize this, assume that simplices in the Vietoris–Rips complex are added one after the other with an associated weight. This gives rise to a *filtration*,

$$\emptyset = \mathrm{K}_0 \subseteq \mathrm{K}_1 \subseteq \cdots \subseteq \mathrm{K}_{n-1} \subseteq \mathrm{K}_n = \mathrm{K},$$

such that each K_i is a valid simplicial subcomplex of K. We write $w(\mathrm{K}_i)$ to denote the weight of K_i.

Similar to what we have previously seen, this gives rise to a sequence of homomorphisms,

$$f_p^{i,j} : H_p(\mathrm{K}_i) \to H_p(\mathrm{K}_j),$$

and a sequence of homology groups, i.e.

$$0 = H_p(\mathrm{K}_0) \xrightarrow{f_p^{0,1}} H_p(\mathrm{K}_1) \xrightarrow{f_p^{1,2}} \ldots \xrightarrow{f_p^{n-2,n-1}} H_p(\mathrm{K}_{n-1}) \xrightarrow{f_p^{n-1,n}} H_p(\mathrm{K}_n) = H_p(\mathrm{K}),$$

where p denotes the dimension of the homology groups.

Persistent homology group

Given two indices $i \leq j$, the p^{th} persistent homology group $H_p^{i,j}$ is defined as

$$H_p^{i,j} := Z_p\left(\mathrm{K}_i\right) / \left(B_p\left(\mathrm{K}_j\right) \cap Z_p\left(\mathrm{K}_i\right)\right),$$

which contains all the homology classes of K_i that are still present in K_j.

We may now *track* the different homology classes through the individual homology groups.

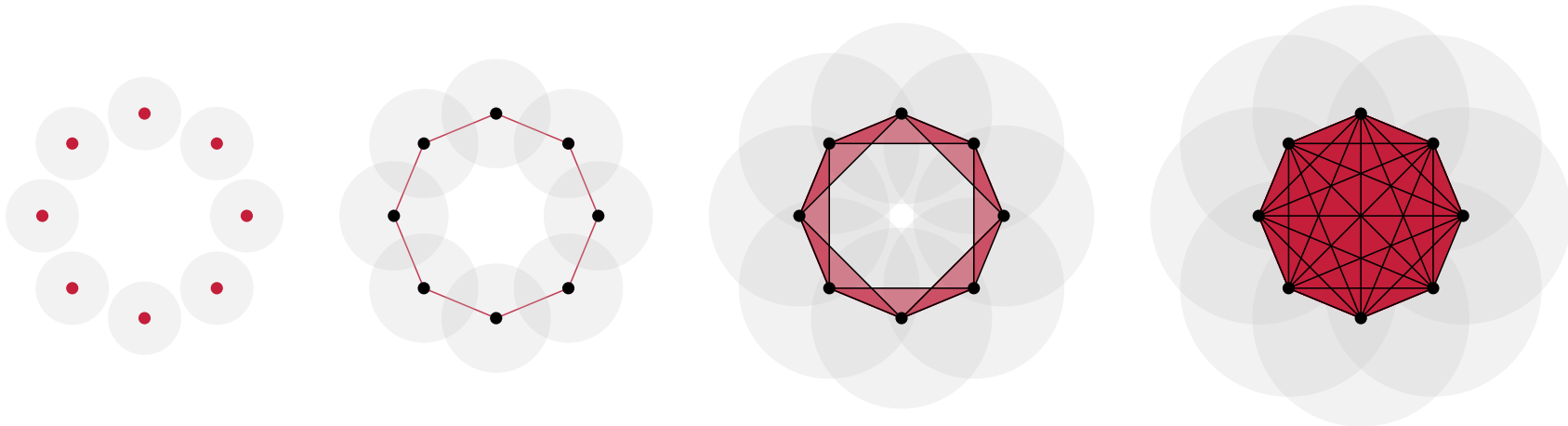
Tracking of homology classes

- *Creation* in K_i : $c \in H_p(\mathrm{K}_i)$, but $c \notin H_p^{i-1,i}$
- *Destruction* in K_j : $f_p^{i,j-1}(c) \notin H_p^{i-1,j-1}$ and $f_p^{i,j}(c) \in H_p^{i-1,j}$

The *persistence* of a class c that is created in K_i and destroyed in K_j is defined as

$$\mathrm{pers}(c) = |w(\mathrm{K}_j) - w(\mathrm{K}_i)|,$$

and measures the 'scale' at which a certain topological feature occurs.



$\epsilon = 0.35$ $\epsilon = 0.53$ $\epsilon = 0.88$ $\epsilon = 1.05$

Here, the topological feature is the circle that underlies the data. It persists from $\epsilon = 0.53$ to $\epsilon = 1.05$, so its persistence is:

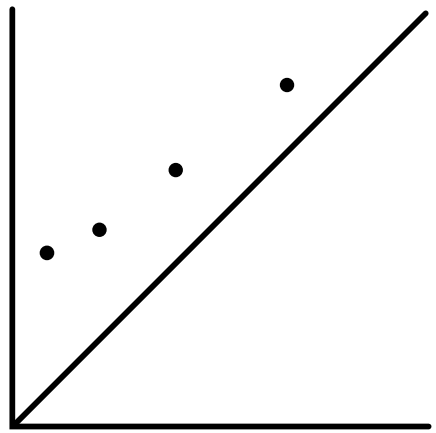
$$\text{pers} = 1.05 - 0.53 = 0.52$$

In general, a high persistence indicates relevant features.

How to represent topological information?

Persistence diagram

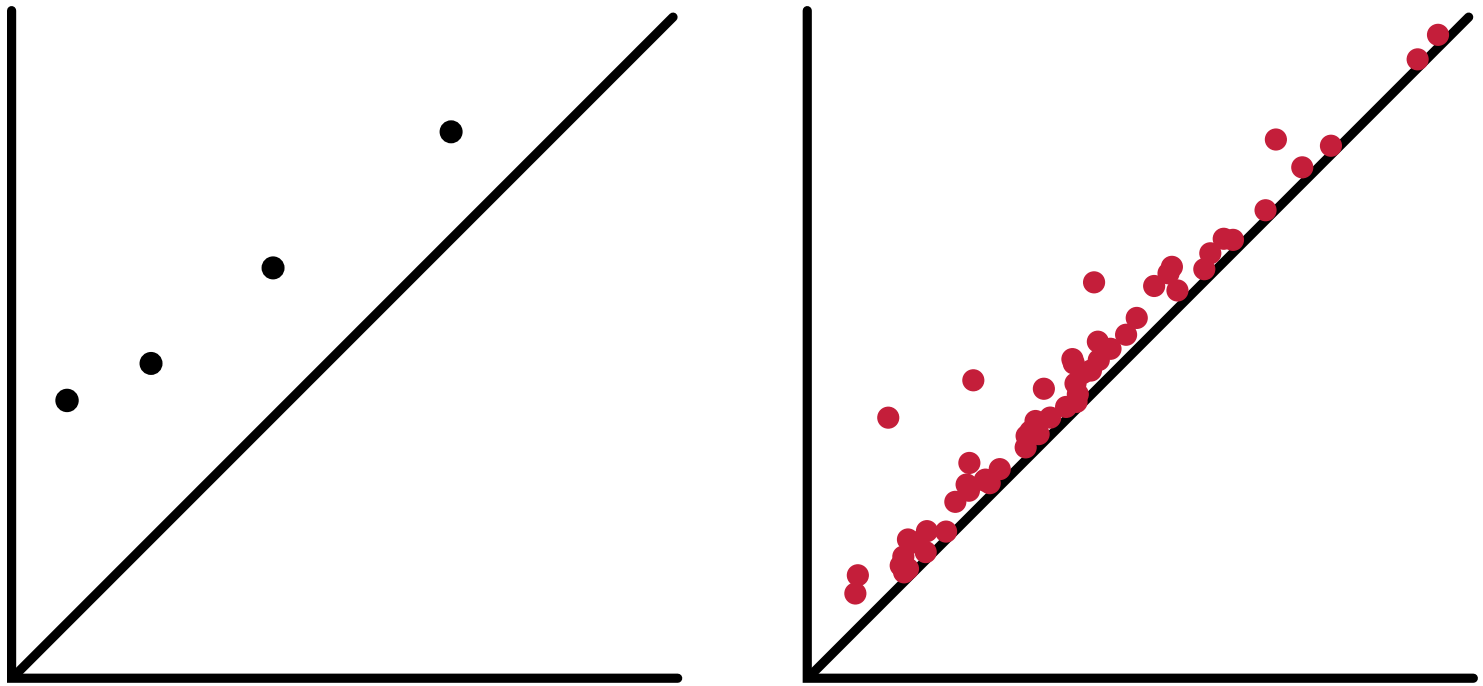
Given a topological feature created in K_i and destroyed in K_j, add a point with coordinates $(w(\mathrm{K}_i), w(\mathrm{K}_j))$ to a diagram:



This *summarizing description* is always two-dimensional, regardless of the dimensionality of the input data!

Uses for persistence diagrams

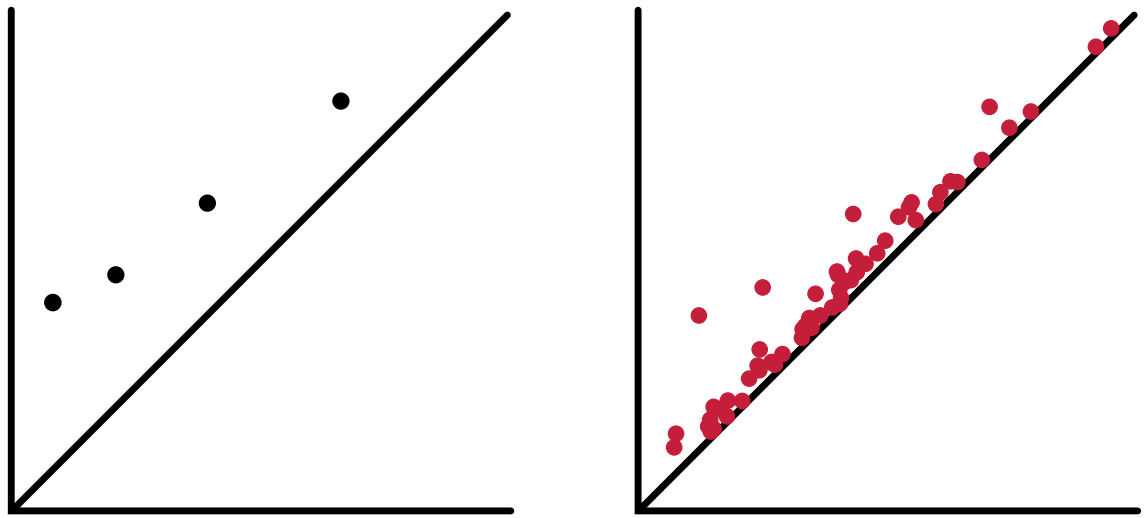
Well-defined distance measures



Persistence diagrams from the same object. Some noise has been added to the object, resulting in spurious topological features. Large-scale features remain the same, though!

Distance measure

Second Wasserstein distance



$$\mathrm{W}_2(X,Y) = \sqrt{\inf_{\eta\colon X \to Y} \sum_{x \in X} \|x - \eta(x)\|_\infty^2}$$

Stability

Theorem

Let f and g be two Lipschitz-continuous functions. There are constants k and C that depend on the input space and on the Lipschitz constants of f and g such that

$$\mathrm{W}_2(X, Y) \leq C \|f - g\|_\infty^{1-\frac{k}{2}}, \tag{1}$$

where X and Y refer to the persistence diagrams of f and g.

Summarizing statistics

Given a *persistence diagram* $\mathcal{D}$, there are various summary statistics that we can calculate:

∞-norm:

$$\|\mathcal{D}\|_\infty = \max_{(x,y)\in\mathcal{D}} |c-d|$$

p-norm:

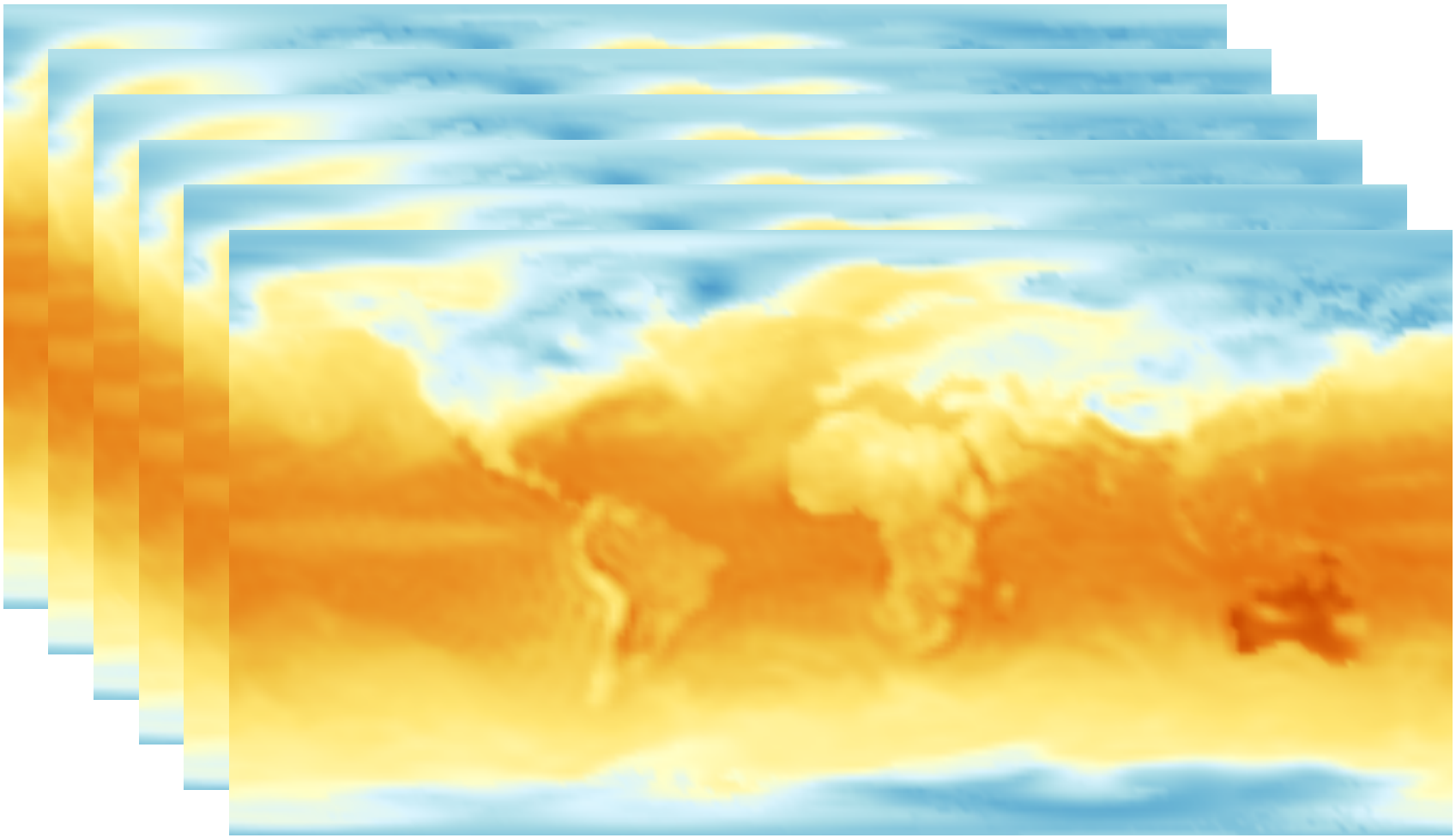
$$\|\mathcal{D}\|_p = \left(\sum_{(x,y)\in\mathcal{D}} (x-y)^p \right)^{\frac{1}{p}}$$

Total persistence:

$$\mathrm{pers}(\mathcal{D})_p = \sum_{(x,y)\in\mathcal{D}} (x-y)^p$$

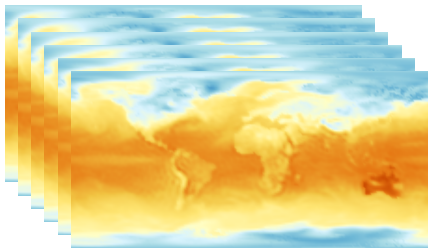
Scalar field analysis

Climate research



Scalar field analysis

What are the issues?

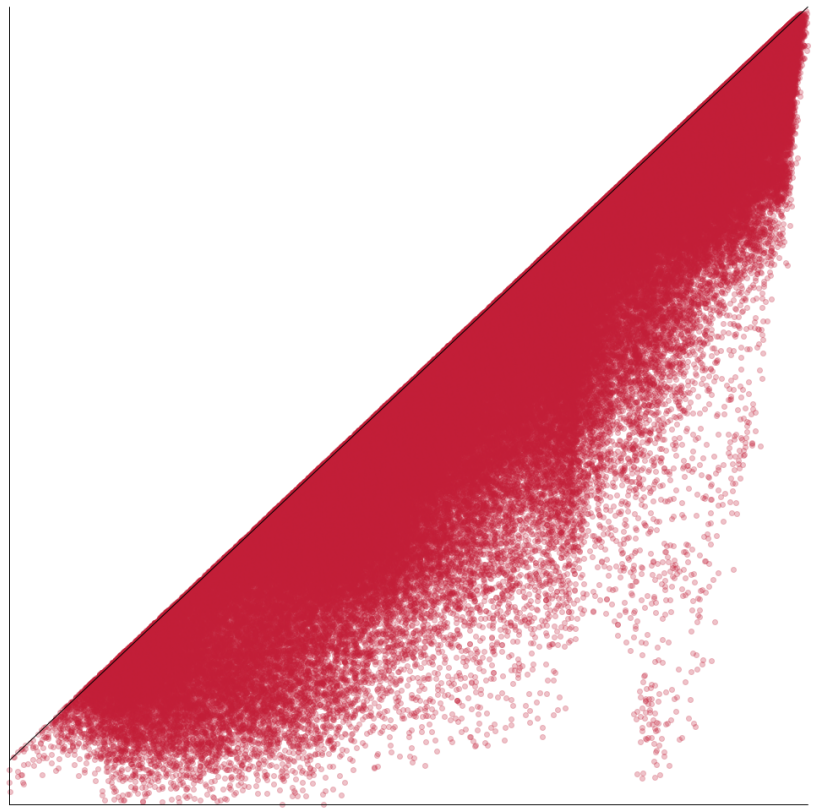


- Need to know about large-scale & small-scale differences in qualitative behaviour of the fields
- *Similar* phenomena may appear at *different* regions in the data
- Time-varying aspects: What are outlying time steps with markedly different properties than the remaining ones?

Using a 2D simplicial complex (surface of the Earth), we can only find topological features in dimensions 0, 1, and 2.

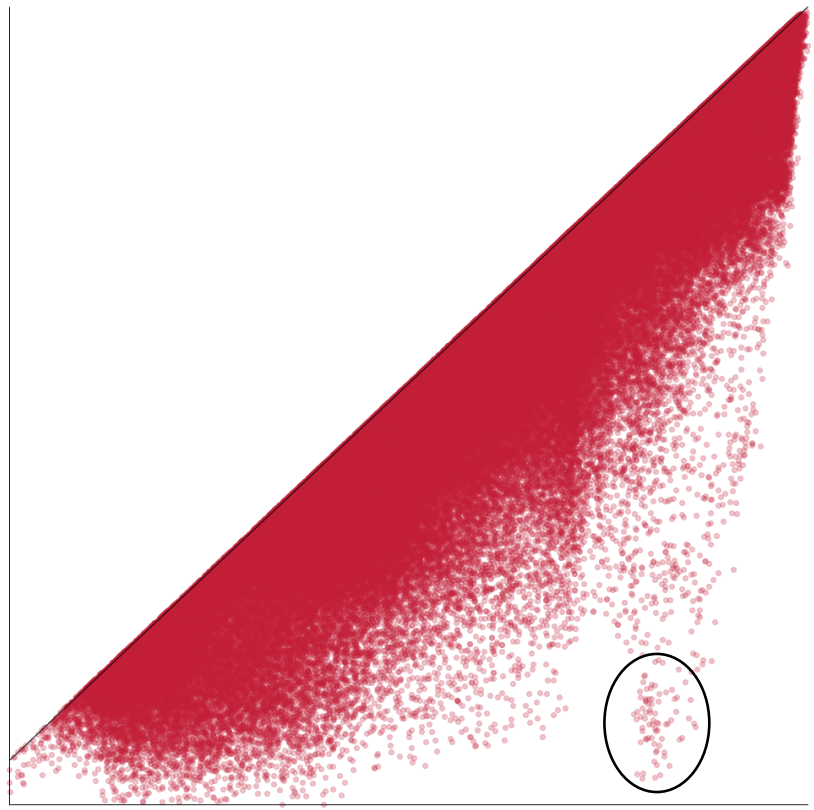
Combined persistence diagram

1460 time steps, dimension 1



Combined persistence diagram

Outliers



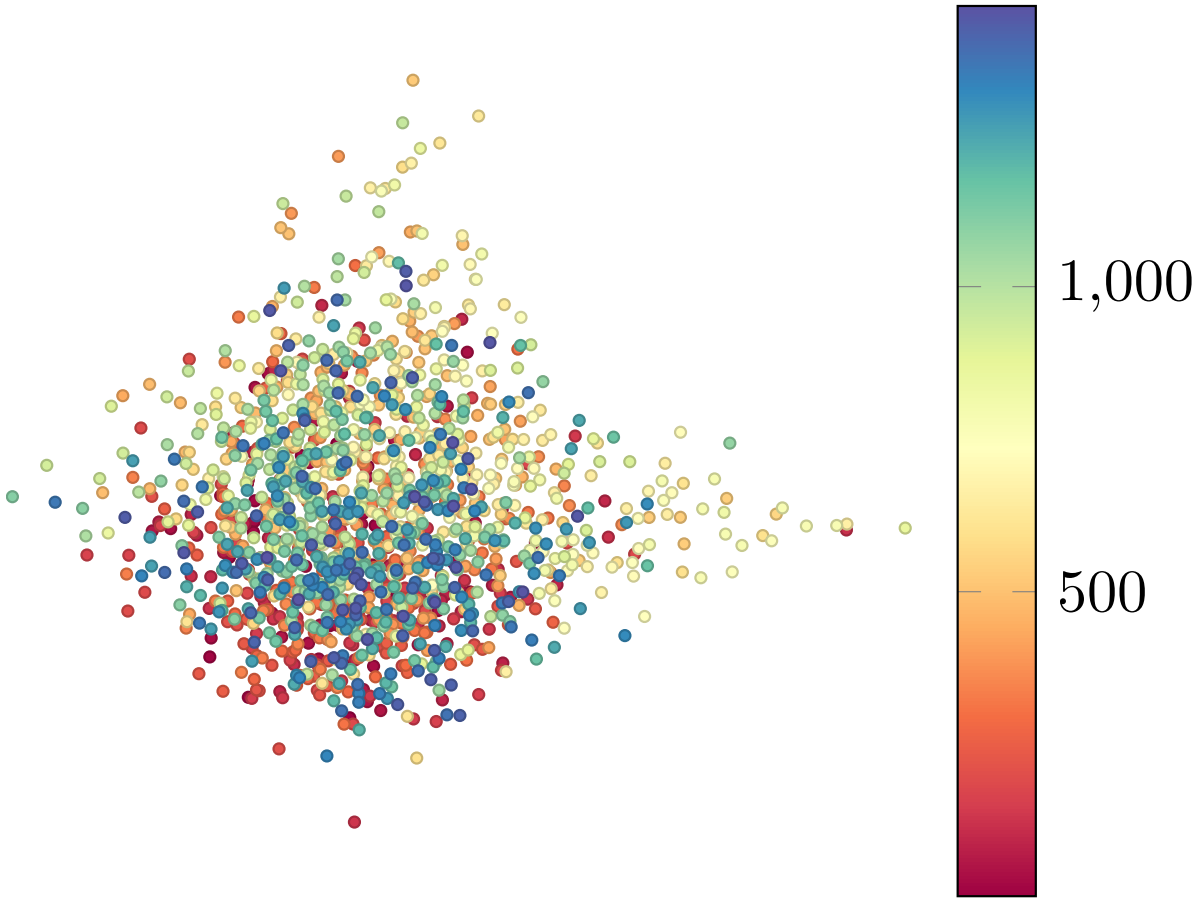
What do the outliers represent?

Time steps in the simulation with extremal temperature phenomena at different places in the world.

Except by visual inspection, this cannot be detected by other methods!

Analysis of cyclical behaviour using summary statistics

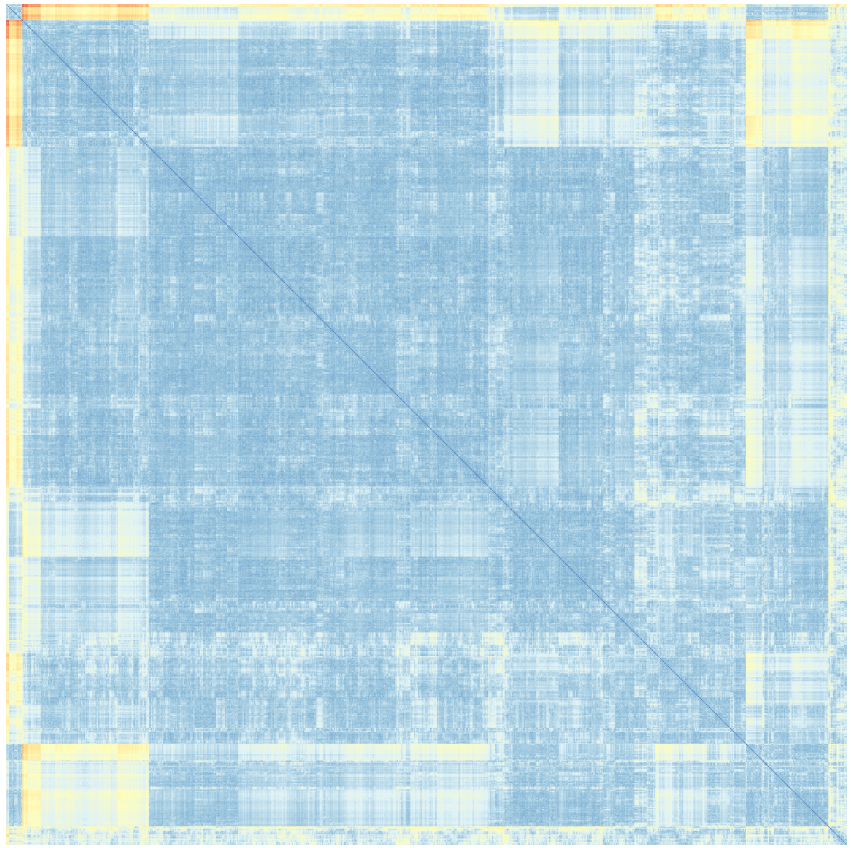
Embedding based on the Wasserstein distance



Outliers can easily be spotted; cyclical behaviour is indicated by points of different colours that are situated next to each other

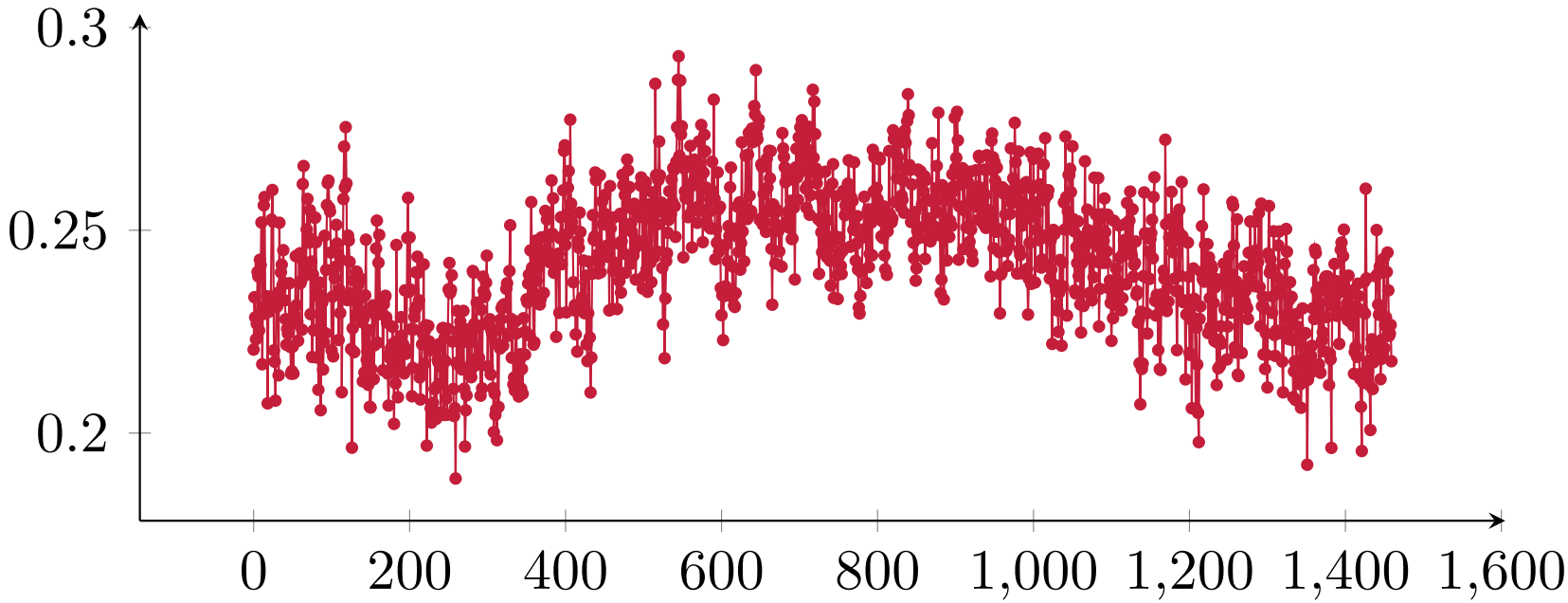
Analysis of cyclical behaviour using summary statistics

Heatmap visualization of the sorted distance matrix



Cyclical structure is hinted at by the block structure.

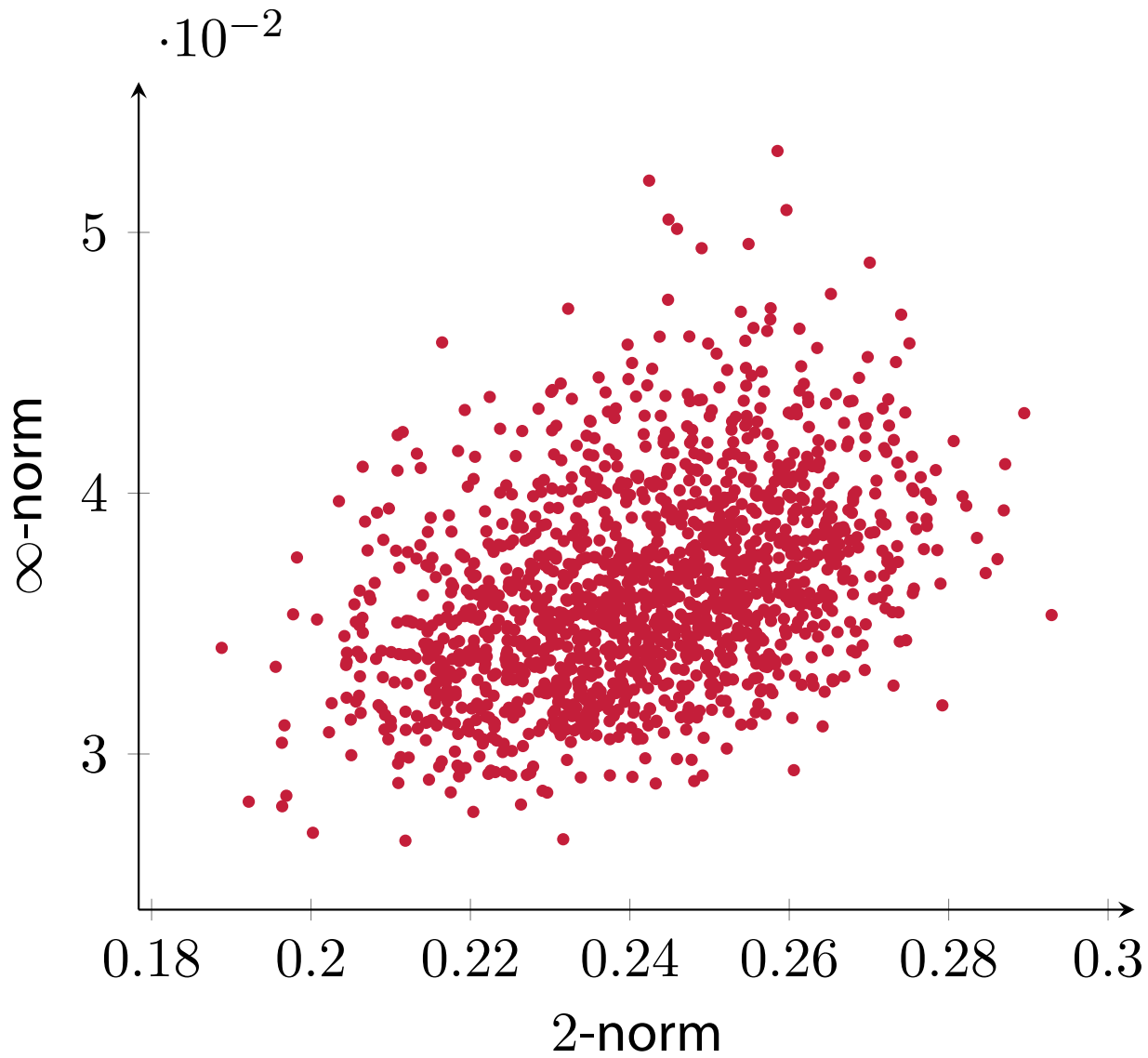
2-norm of all persistence diagrams



Detection of cyclical behaviour (seasons, micro-climate) regardless of the physical location.

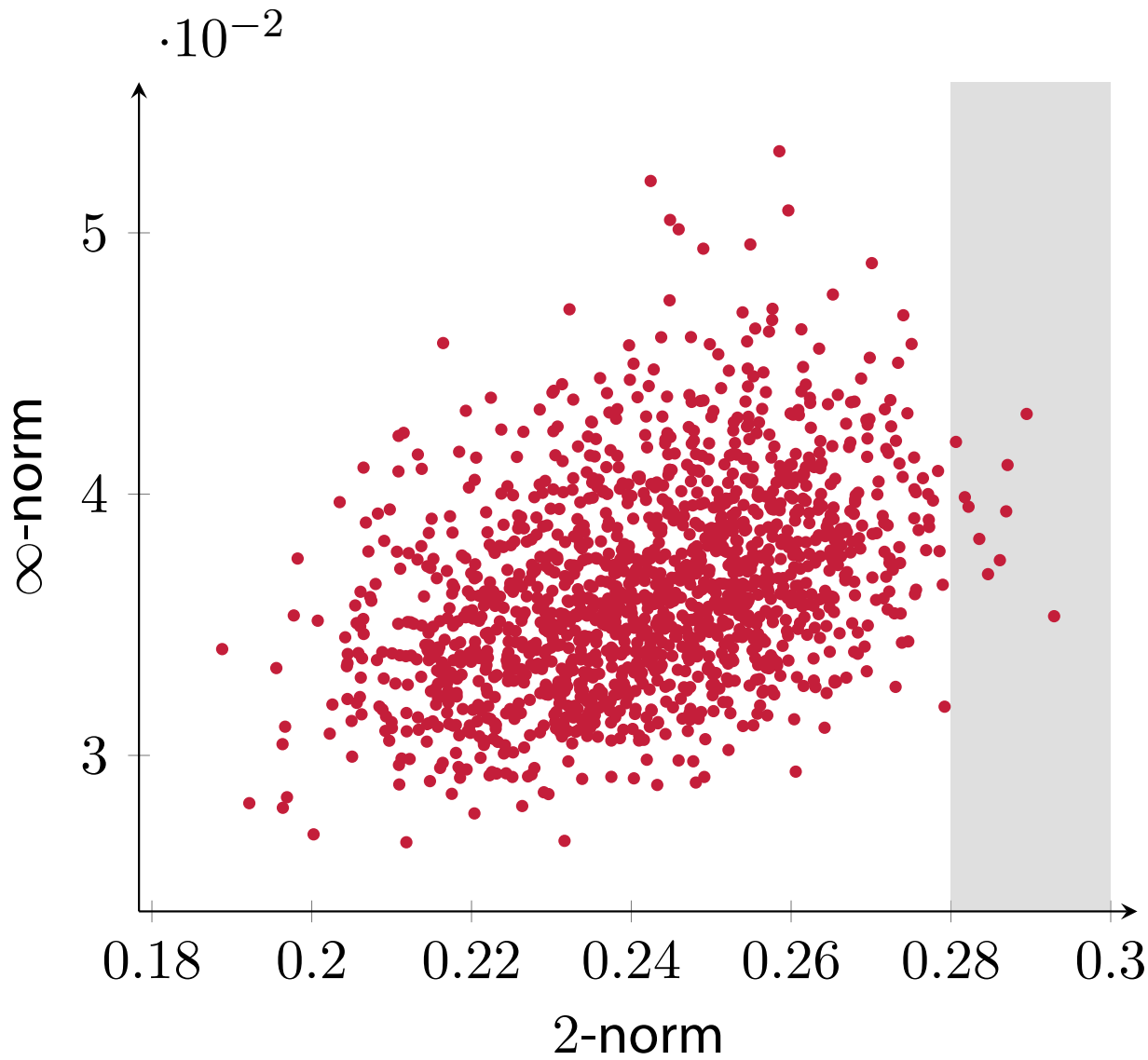
2-norm vs. ∞-norm

All time steps



2-norm vs. ∞-norm

Interesting time steps: Large 2-norm, small ∞-norm



Conclusion

Take-away messages

1. Persistent homology is a new way of looking at complex data.
2. It has a rich mathematical theory and many desirable properties (robustness, invariance).
3. Lots of interesting applications.

Interested? Drop me a line at bastian.rieck@iwr.uni-heidelberg.de!