# Multivariate Data Analysis Using Persistence-Based Filtering and Topological Signatures

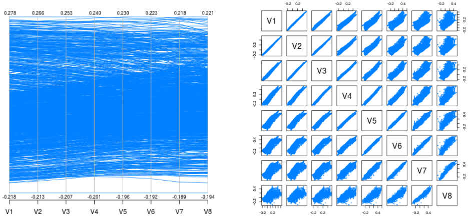
**Bastian Rieck** Hubert Mara Heike Leitte

Research group *Computer Graphics and Visualization*
Interdisciplinary Center for Scientific Computing
Heidelberg University, Germany

October 18, 2012

Universität Heidelberg
Zukunft. Seit 1386.

# Motivation




(created with R)

## Setting

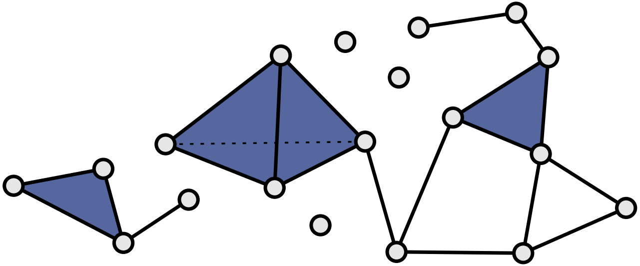
- High-dimensional ($\gg 4$) scientific data
- Understanding the *shape of data*
- Our approach: Algebraic topology

# Simplicial complex

## The basic building block of algebraic topology

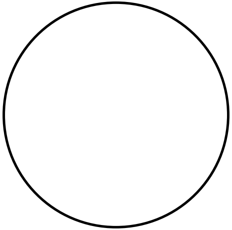
A simplicial complex consists of:

- 0-simplices (vertices)
- 1-simplices (edges)
- 2-simplices (triangles)
- 3-simplices (tetrahedra)
- ...

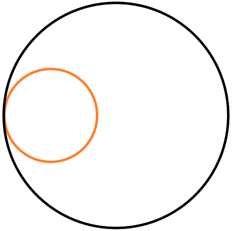


# Homology groups

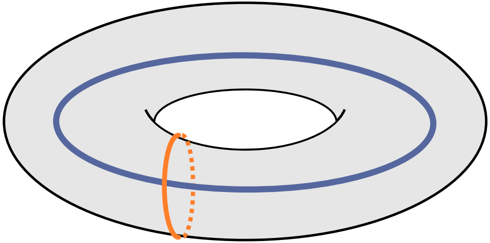
- One group per dimension
- Rank of $k$th group = number of $k$-dimensional holes = $b_k$
  - Connected components
  - Loops
  - Tunnels (voids)
  - ...



$b_0 = 1,\ b_1 = 1,\ b_2 = 0$



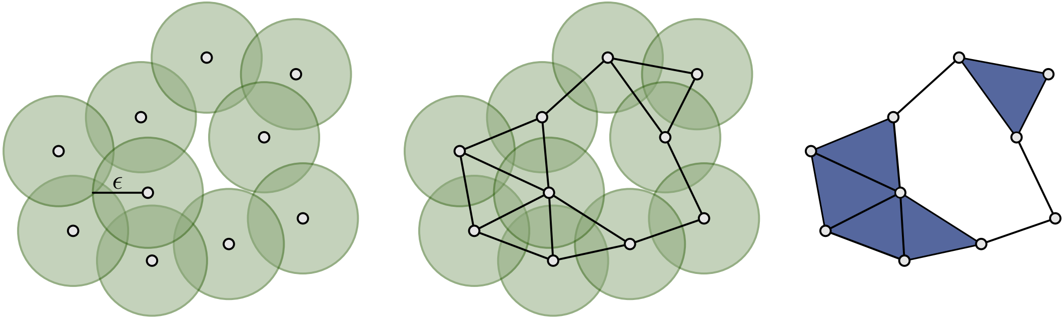
$b_0 = 1,\ b_1 = 2,\ b_2 = 0$



$b_0 = 1,\ b_1 = 2,\ b_2 = 1$

# Topological recipe for scientific data

- Goal: "Convert" input data to simplicial complex
- Requires: Distance function on input data (Euclidean distance, $p$-norm, ...) and distance threshold parameter $\epsilon$
- Use $\epsilon$ to obtain neighbourhood graph
- *Expand* graph to simplicial complex

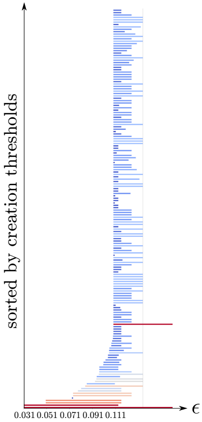


## Persistent homology calculation for simplicial complex

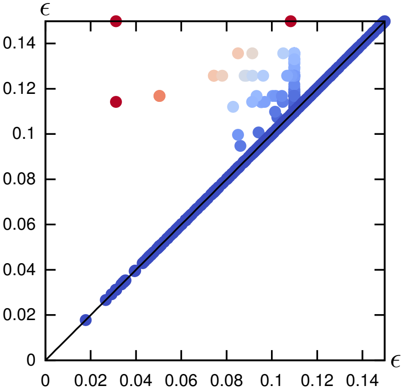
- Obtain homology groups for all values $\leq \epsilon$
- Each $k$-dimensional hole is represented by $a, b \in \mathbb{R} \cup \{\infty\}$
- $a$: threshold at which $k$-dimensional hole is created
- $b$: threshold at which $k$-dimensional hole is destroyed
- Persistence $:= b - a$
- The *larger* the persistence, the more *important* the feature!

# How to visualize persistence intervals of a given dimension?

- $(a, b)$: pair of creator-destroyer thresholds
- coloured by persistence value

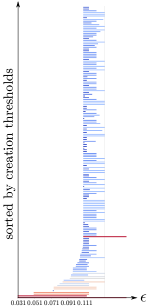


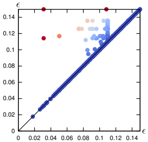

Draw a line from $a$ to $b$ for each pair $(a, b)$

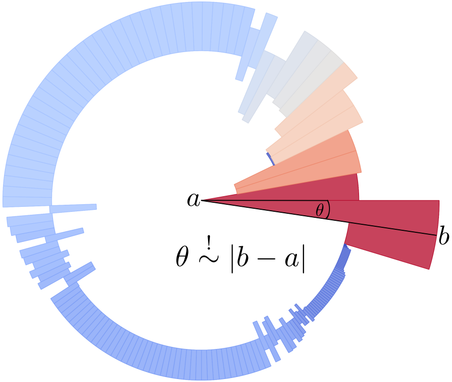



Draw point $(a, b)$ for each pair $(a, b)$

# Persistence rings — an alternative visualization




Allocate an annular segment from radius $a$ to radius $b$ for each pair $(a, b)$ in dimension $k$

# Our workflow for high-dimensional data sets

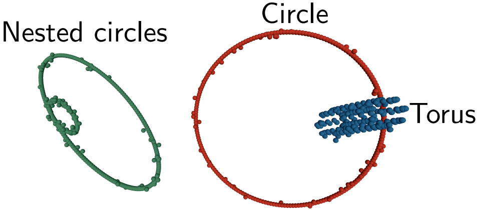
## Philosophy

**clustering + topological signatures ⇒ improved understanding**

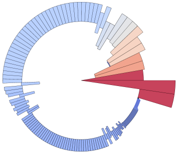
1. Accept generic point clouds as input
2. Use persistence-based clustering scheme of Chazal et al.
3. Obtain a topologically-based clustering of the data set
4. Calculate topological signatures for each cluster: Apply persistent homology algorithm

# Results for synthetic data

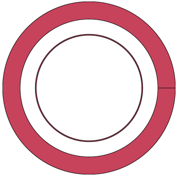
- $n = 1000$
- $d = 60$
- $t = 3\mathrm{s}$



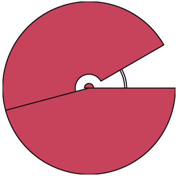

Projection of synthetic data set



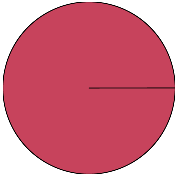
Torus (dim. 1)



Torus (dim. 2)



Nested circles (dim. 1)

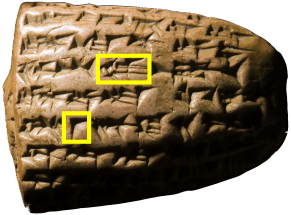


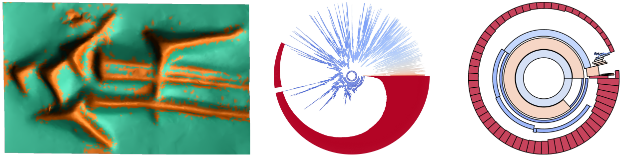
Circle (dim. 1)

# Results for cultural heritage data

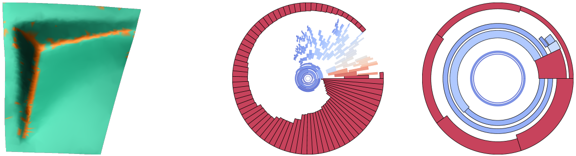
## Noisy input data

- MSII curvature estimation
- $n = 1000\text{–}15000$
- $d = 16$
- $t = 5\mathrm{s}\text{–}10\mathrm{s}$





Background Writing



Background Writing

## Summary

- Analysis of high-dimensional data sets using algebraic topology
- *Persistence rings* as a new visualization metaphor
- Structural description of data set (for every dimension)
- Applicable to data sets of arbitrary dimensions

# Thank you for listening!

## Acknowledgements

Research group Computer Graphics and Visualization (**CoVis**)
Interdisciplinary Center for Scientific Computing
Heidelberg University

HGS MathComp

Universität Heidelberg
Zukunft. Seit 1386.

IWR

E-mail: `bastian.rieck@iwr.uni-heidelberg.de`
URL: `http://www.iwr.uni-heidelberg.de/groups/CoVis`