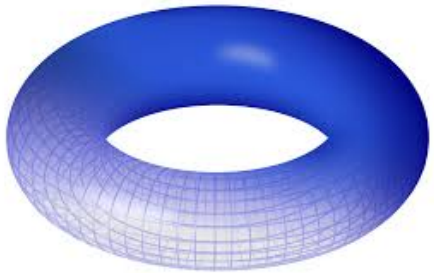
Topological Data Analysis

Robbert Fokkink

r.j.fokkink@tudelft.nl

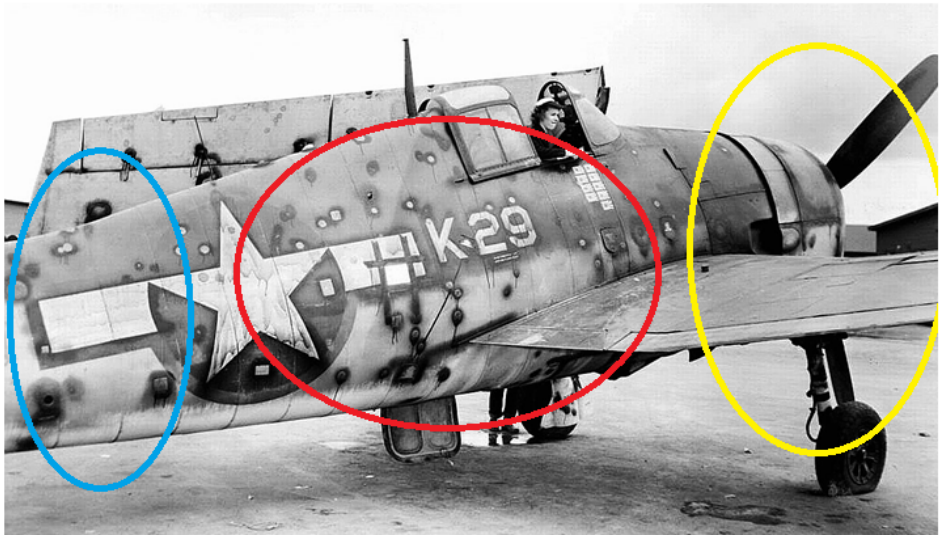
About this lecture



- TDA applies to: clustering, pattern recognition.
- A new method for you. This is an exercise in quickly getting acquainted with a new tool. A useful skill for your project. TDA itself is not necessarily useful in your project.
- The T in TDA stands for topology. Do you know what topology is?

Stats Basics

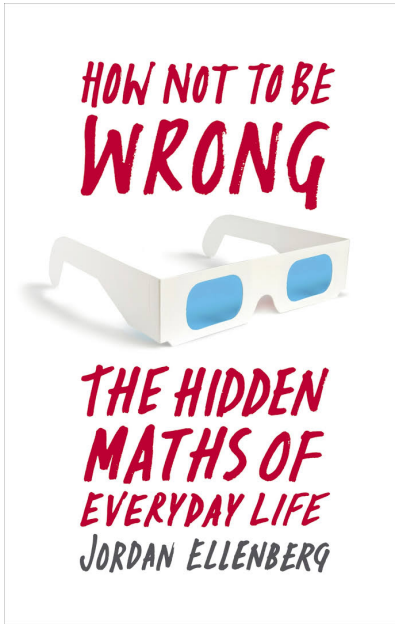
Where would you fortify the plane?



A rear B middle C engine

Stats Basics

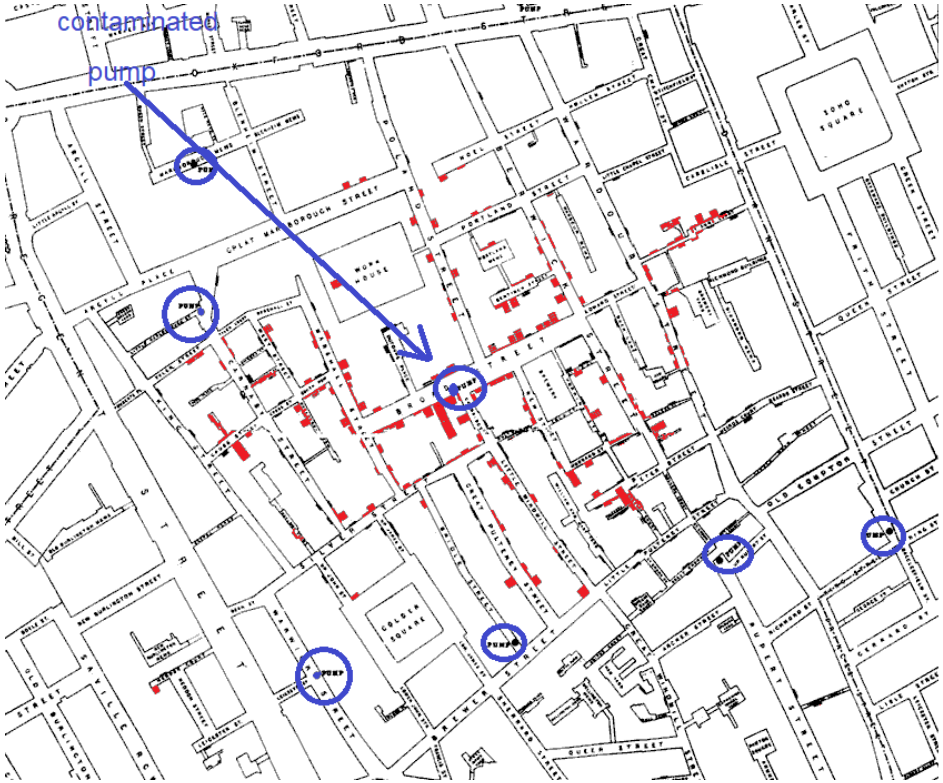
Thanks to covid, everybody's got plenty of time. Please read



or watch the video. Amazon.com writes: *A math-world superstar unveils the hidden beauty and logic of the world and puts its power in our hands* I would say: a top mathematician explains statistics.

Stats Basics

Our eyes are powerful data analysts.
John Snow located the origin of the 1854 London cholera outbreak to a water pump in Broad Street by using a map:



Thanks to the computer, we can even animate the data. A famous animation is by Hans Rosling

Topological Data Analysis converts data into a barcode

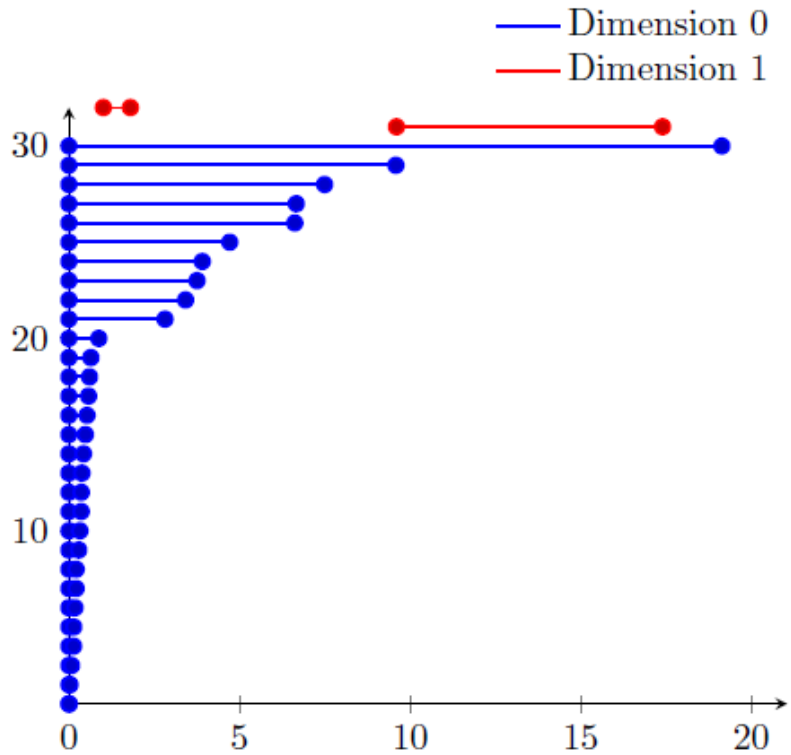


Figure: A persistance graph for dimensions 0 and 1. Source: calculation of Hawkeye data – Esther Visser, MSc thesis 2018.

Today's lecture

- What is topology?
- What is a persistence diagram?
- How do you convert diagrams into statistics?
- What is the assignment?

Topology

Topology is the study of shapes.

Dutch pride: topology got started in 1631 by a Frenchman who stayed in an Amsterdam hostel, which was later replaced by a city hall, which was turned into a palace by another Frenchman.

It was there that René Descartes invented angular defect and noticed how it can be used to say something about shapes.

Topology

The three figures below all have the shape of a ball. The figures are mostly flat except at the corner points. The defect at a corner is:

$$2\pi \ - \ \text{sum of interior angles}$$

- The cube has 8 corners. Defect at each corner $\frac{\pi}{2}$. Total defect 4π.
- The dodecahedron has 20 corners. Defects $\frac{\pi}{5}$. Total defect 4π.
- The octahedron has 6 corners. Defects $\frac{2\pi}{3}$. Total defect 4π.

Theorem (Descartes)

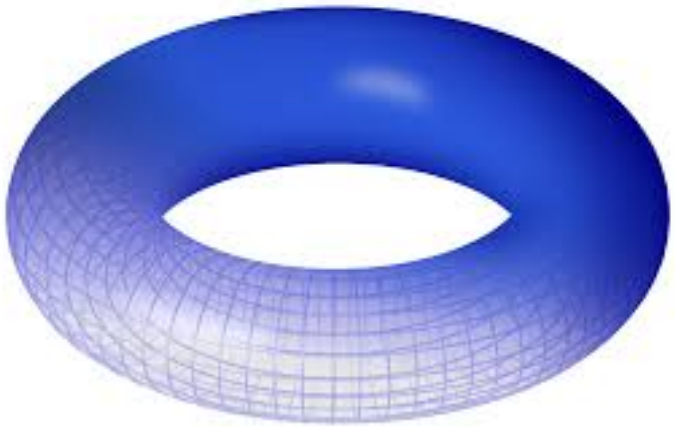
For a polyhedron with V vertices, E edges and F faces, the sum of the defects equals

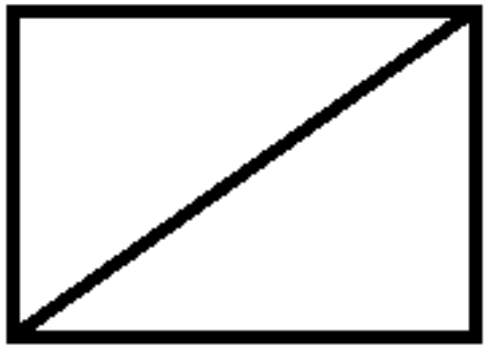
$$2\pi\left(V - E + F\right)$$

The alternating sum $V - E + F$ is now known as the Euler number. All three polyhedra have the same shape (that of a sphere) and the Euler number captures it.

Topology

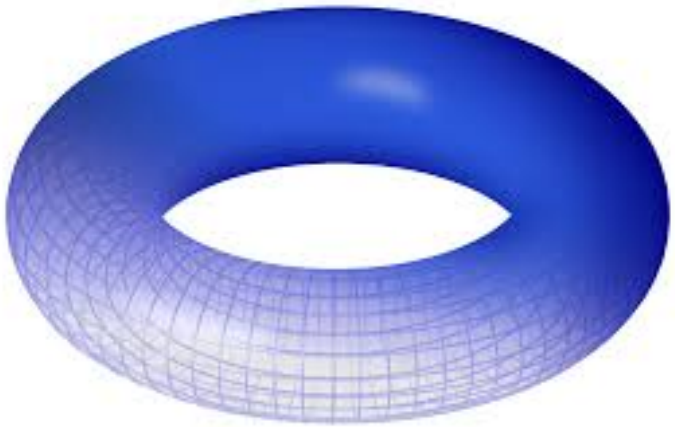
What is the Euler number of a torus?

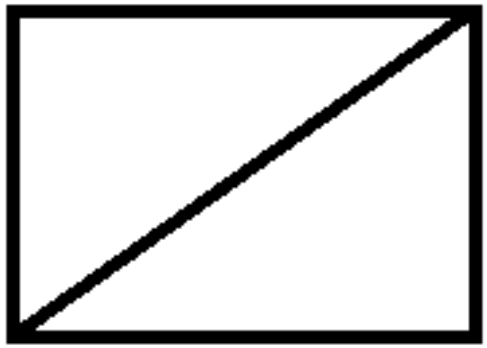




Topology

What is the Euler number of a torus?



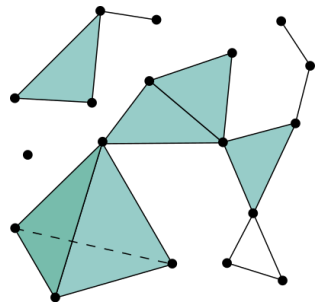


$$F - 3F/2 + 3F/6$$

Topology

Moving on to Pisa in 1871, where Enrico Betti defined Betti numbers. Note the plural here: we have a sequence of numbers. The k-th Betti number counts the number of $k+1$-dimensional holes.

A polyhedron consists of vertices, edges, faces, and since we are mathematicians we continue just as easily in higher dimensions. This object is called a simplicial complex.



Betti defined a number for each dimension, starting from dimension zero. The numbers for the figure are: $b_0 = 3$[1], $b_1 = 1$, $b_2 = 1$. Three connected components, one loop, one hole.

[1] Some people say $b_0 = 2$, the number of connected components minus one

Topology

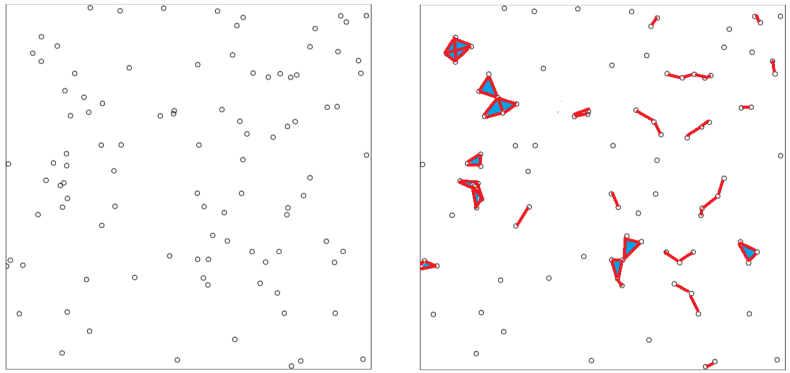
The precise definition of the Betti numbers of a simplicial complex is a bit technical.

Define vector spaces $V_0, V_1, V_2, \ldots$ by $V_k = \mathbb{R}^{k-\text{simplices}}$ and define a map δ that maps a k-simplex onto its boundary. This gives a chain of maps $\delta\colon V_{k+1} \to V_k$ that has the property that $\delta^2 = 0$.

The Betti number expresses the defect between kernel and image of δ. Just like Descartes' concept of defect of an angle.

Persistence diagrams

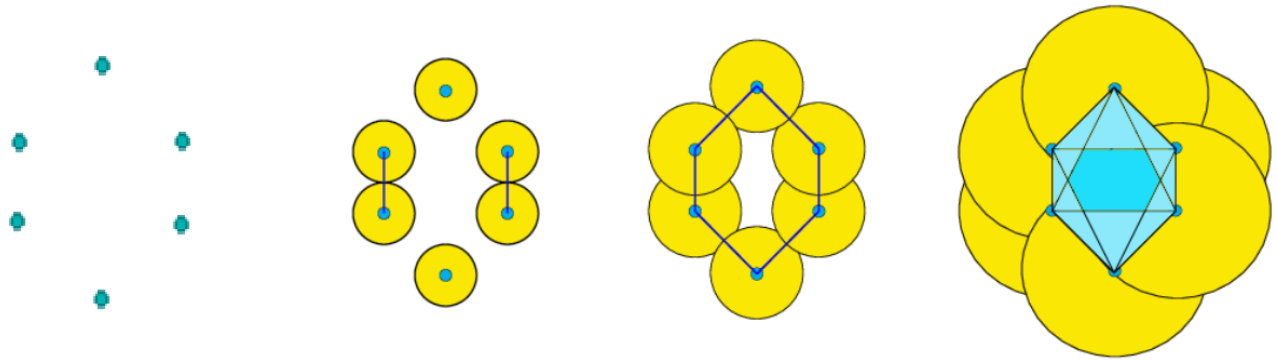
A data set is just a scatter of points. How can topology be useful? We move on to Vienna, where Leopold Vietoris in 1927 defined what is now known as the Vietoris-Rips complex, or simply Rips complex.



On the left you see the data scatter. On the right you see a simplicial complex: if points x and y are close enough, connect by an edge. If x, y, z are close, connect by a face. If x, y, z, w are close, connect by a simplex. Etc. Now we have a simplicial complex and we can compute its Betti numbers.

Persistence diagrams

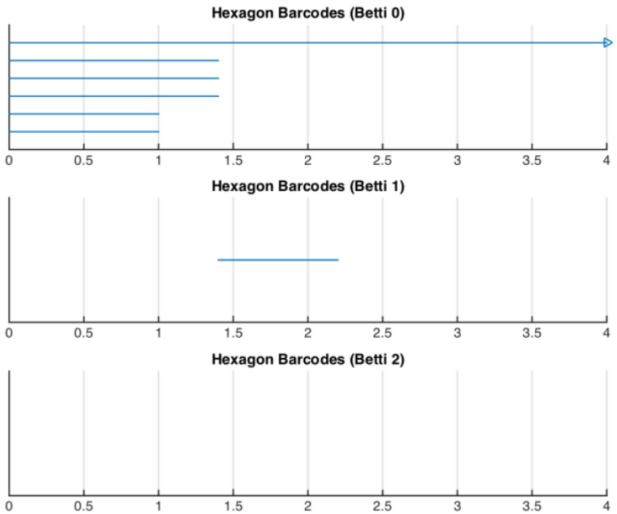
The Rips complex depends on what we call close.



The yellow balls indicate what we call close. All Rips complexes have 6 points, that's our data: a hexagon. As the yellow balls grow, edges and faces appear in the complex. The zero-th Betti number for the four complexes is $b_0 = 6, 4, 1, 1$. The first Betti number is $b_1 = 0, 0, 1, 0$. All higher Betti numbers are zero.

Persistence diagrams

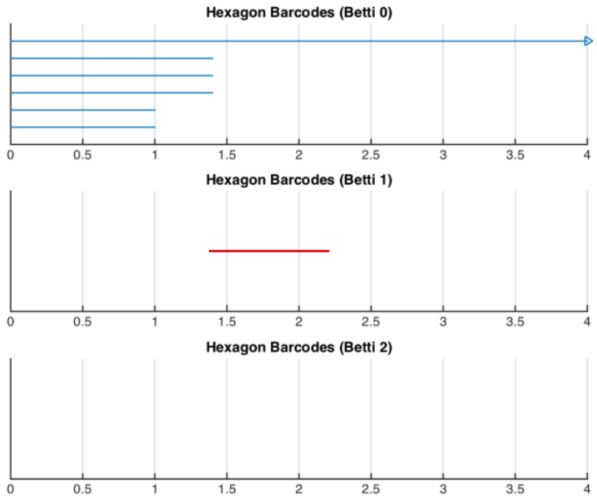
The Betti number depends on the radius of the balls. A barcode plots the Betti numbers versus the radius.

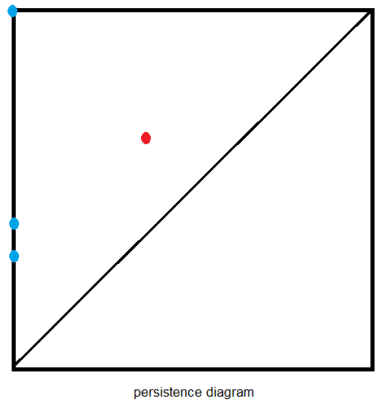


As we saw b_0 goes down from 6 to 4 to 1. And b_1 goes up from 0 to 1 and then back down to 0 again. Finally, b_2 is always 0. As you can see in the three barcodes for b_0, b_1, b_2. Hans Rosling would animate a barcode, starting with 6 points, moving closer and merging until finally one points remains.

Persistence diagrams

Each bar has a beginning a and an end b. It can be plotted more succinctly as a point (a, b). If we collect all these points, then we get the persistence diagram:

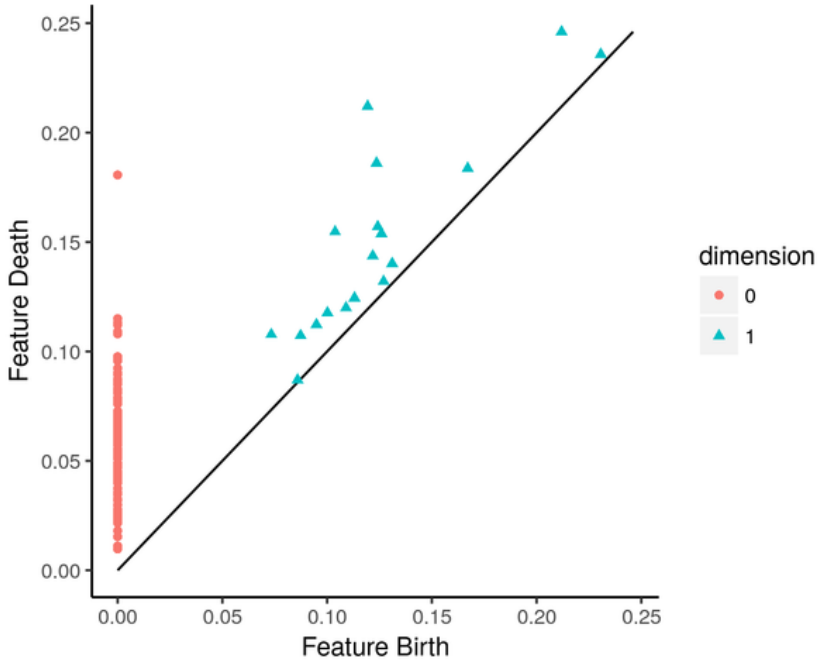




In this example, the single red point for Betti number b_1 stands out in the persistence diagram. It tells us that the shape is a single loop. A hexagon, which can be visualized by the Rips complex using a distance of around 1.75.

Persistence diagrams

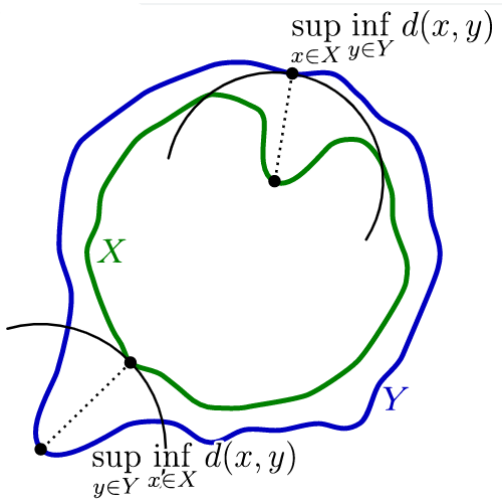
For big data sets, the diagram will contain many more points



This diagram suggests that a good distance to visualize the Rips complex is around 0.15, when the complex has one connected component and five loops.

Statistics

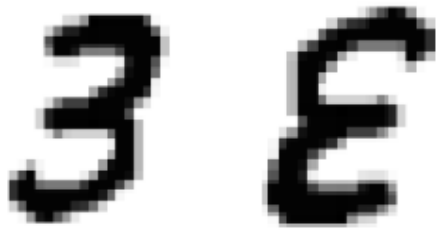
Statisticians need statistics, i.e., numbers. TDA produces persistence diagrams. What good is that? We go back to Greifswald, a small port at the Baltic Sea, where Felix Hausdorff in 1914 defined Hausdorff distance.



The distance $d(x, Y)$ between a point and a closed set is defined as $\min d(x, y)$ for all $y \in Y$. Let $x \in X$ be the point that is furthest away from Y and let $y \in Y$ be the point that is furthest away from X. The Hausdorff distance $d(X, Y)$ is the maximum of $d(x, Y)$ and $d(X, y)$.

Statistics

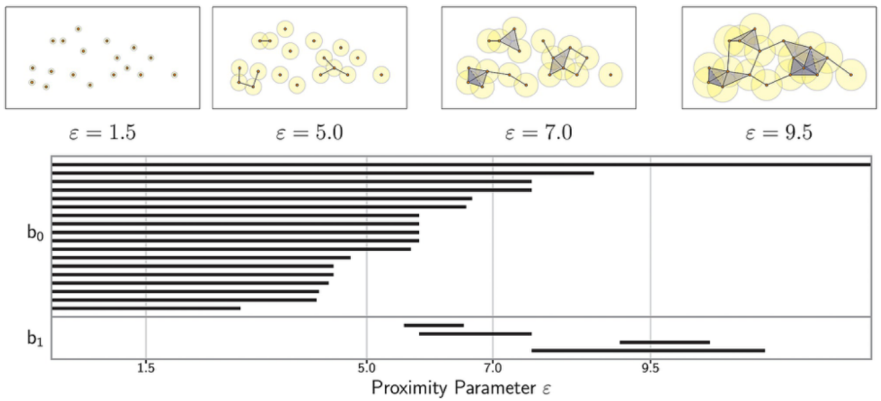
Therefore, it is possible to define the distance between two persistence diagrams. In TDA this is called the bottleneck distance.



Hypothesis test: are these two shapes the same? Generate 100 samples of the shape on the left, for instance by bootstrapping, or by adding random noise, or by random geometric transformations. Compute the bottleneck distances and compare these to the distance between the figures on the left and the right.

Statistics

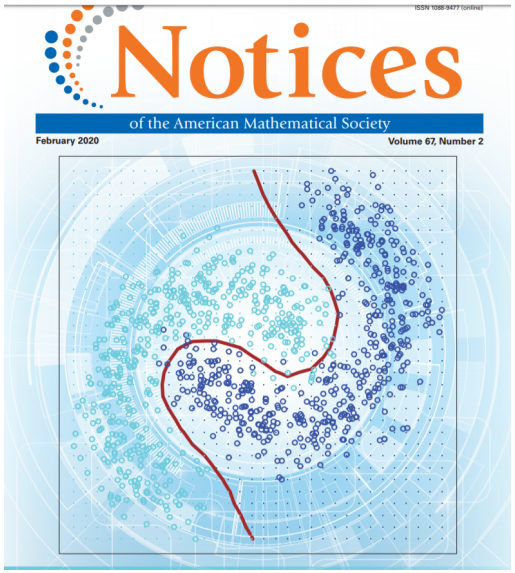
TDA is related to clustering. The relation is so close that Gunnar Carlsson, the inventor of TDA, explains in a post that TDA is not the same as clustering. In your assignment, you are going to use TDA for clustering anyway.



The barcodes suggest that we may cluster at $\epsilon = 7$, when there are four clusters, or at $\epsilon = 9.5$, when there is one cluster with two loops.

Assignment

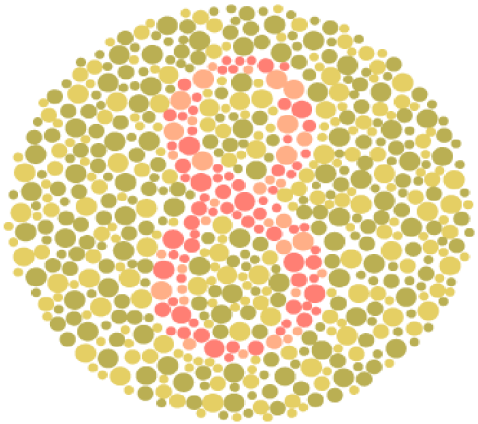
Your first task is to reproduce a clustering that is on the front cover of the Notices of the AMS



It shows a brown dividing line between two clusters, computed by a machine learning algorithm. Your task is to compute the barcode and see what distance corresponds to two (or three or four) connected components for the Rips complex. You do not have to draw the brown dividing line.

Assignment

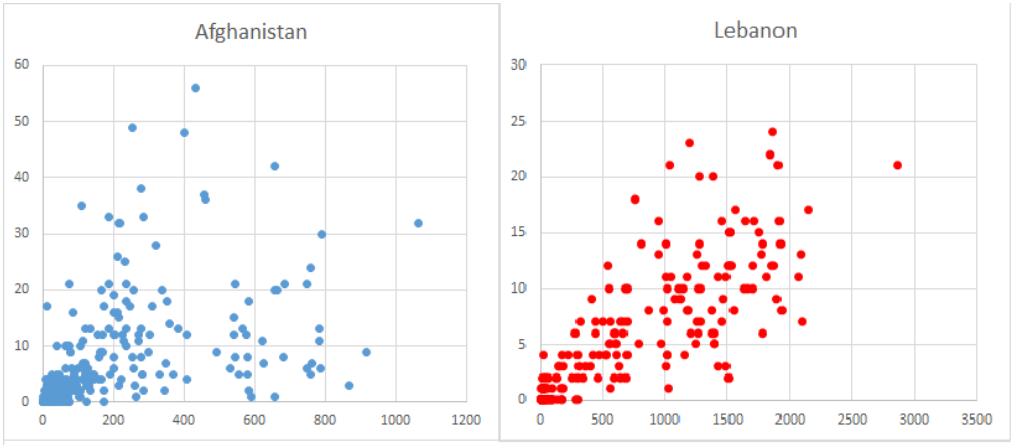
Your second task is a color blindness test for TDA



You get a data set that represents a number. It is not difficult for you to visualize that number, but can TDA do it? This is another clustering exercise. Find the right ϵ so that the Rips complex falls apart into connected components that show the number.

Assignment

Your third task has to do with Covid19. Here you see a scatter of daily cases and deaths in Afghanistan and in Lebanon.



daily number of cases and deaths during 2020

How close are these scatters according to TDA? Compute the barcodes, determine the bottleneck distance. Take your ten favorite countries from the data set, determine which countries are closest according to the bottleneck distance.

Assignment

- Form groups of 2 to 5.
- Find TDA software on the net. You can find tools in R or Python.
- Analyse the data sets on Brightspace that belong to the three tasks.
- Write a short report – max 6 pages – and upload it before March 8th, 24.00 hrs.