Topological Data Analysis: Persistent Homology

Dan Christensen
University of Western Ontario

April 13, 2018

Figures and images are all due to the people cited.

Recall

For a finite metric space C and $\epsilon \geq 0$, the Vietoris-Rips complex $VR_\epsilon(C)$ is the abstract delta complex with n-simplices $\{x_0 < x_1 < \cdots < x_n \mid d(x_i, x_j) \leq \epsilon \; \forall i, j\}$ and where d_i^n omits x_i.

Here C has been equipped with an arbitrary total order.

Theorem. Suppose $V_0 \to V_1 \to V_2 \to \cdots$ is a sequence of finite dimensional vector spaces which stabilizes. Then the sequence is isomorphic to a direct sum of $\mathbb{F}[a, b]$'s, unique up to order.

The big picture:

$$\begin{aligned} \{\text{finite metric spaces}\} &\xrightarrow{VR} \{\epsilon\text{-indexed delta complexes}\} \\ &\xrightarrow{H_*} \{\text{an } \epsilon\text{-indexed vector space for each } n\} \\ &\longrightarrow \{\text{a barcode for each } n\} \\ &\longrightarrow \{\text{enlightenment}\} \end{aligned}$$

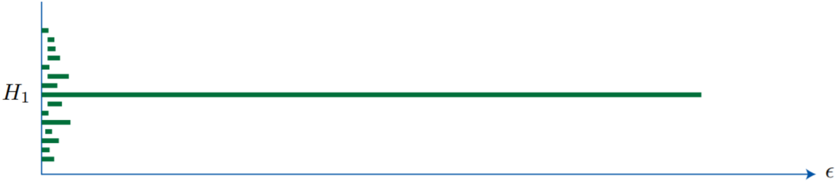
The enlightment will hopefully come with the interactive demo I show now.

Natural images

Mumford, Lee, and Pederson took a collection of 4167 outdoor digital photographs, sampled random 3x3 regions in these images, and kept those with the most contrast, ending up with 8,000,000 points in $\mathbb{R}^9$.

They did some further processing, and also filtered the data to the densest part, ending up with 5,000 points.

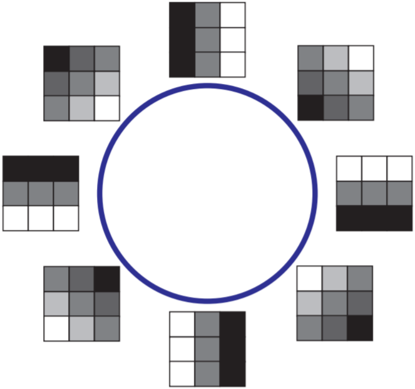
The resulting H_1 barcode is:



So there is a clear circular structure to the data.

Natural images

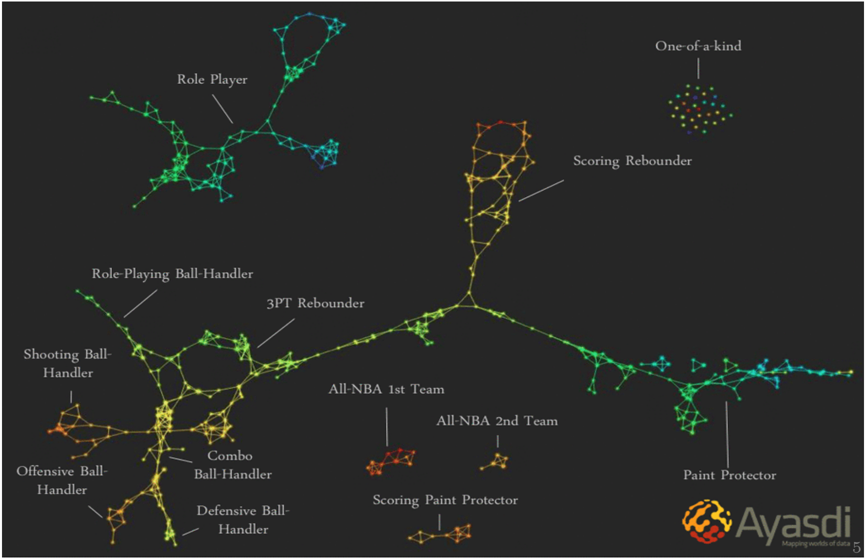
Further work revealed that the circle has to do with the orientation of the gradient direction:

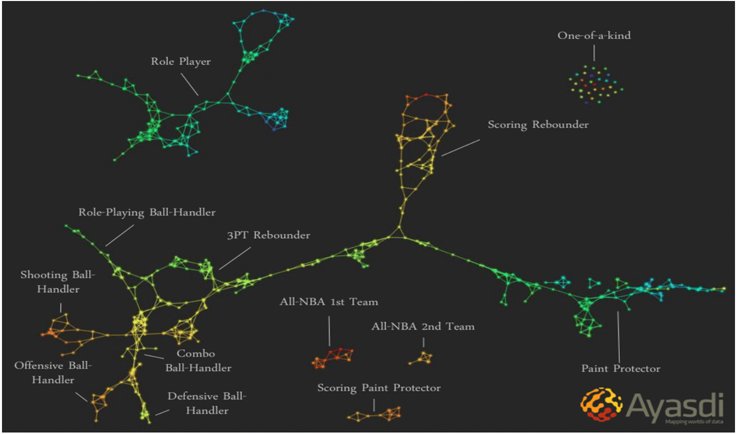


Even more, different filtering revealed two more circles, and eventually to the insight that the Klein bottle appears in the data. (The full story could be a talk in our seminar.)

Basketball

Muthu Alagappan analyzed stats for 452 NBA players using a tool to cluster and link the data using TDA:





He found that players naturally grouped into 13 new “positions” that more accurately reflect playing styles and skills.

He has slides that illustrate how he thinks this will allow teams to improve their results.

Another possible talk, including a discussion of the clustering technique.

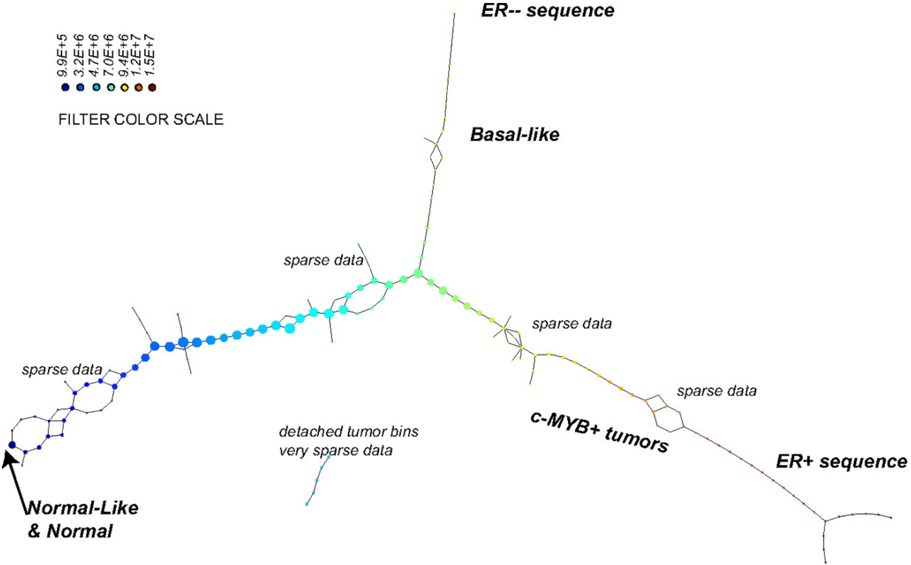
Breast cancer

Nicolau, Levine and Carlsson studied existing data on breast cancer tumors from 495 patients.

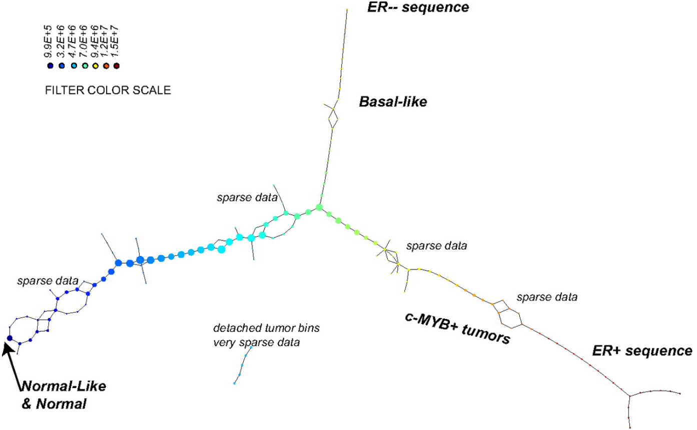
Each tumor has a "gene expression profile," giving a point in $\mathbb{R}^{24,479}$.

The data was massaged down to $\mathbb{R}^{262}$, and then a TDA clustering tool called Mapper was used to group the data into a network.

Breast cancer



Breast cancer



The section labelled “c-MYB+” is a newly identified form of breast cancer which is not fatal, and makes up 7.5% of the cases.

Standard tools don’t see this cluster.

References

- Ghrist, Barcodes: the persistent topology of data
- Ghrist, Homological algebra and data
- Carlsson, Topology and data
- Carlsson, Topological pattern recognition for point cloud data
- Edelsbrunner and Harer, Persistent homology — a survey
- Chazal and Michel, An introduction to TDA: fundamental and practical aspects for data scientists
- And lots more…

Onwards

There are many more applications, and lots of theoretical tools we could talk about in the learning seminar, including:

- The Čech complex, how it compares.
- Density sampling.
- Clustering.
- More case studies.
- The Mapper algorithm, which is used by many of the most interesting applications.
- Sublevel set filtration, given $f : \{data\} \to \mathbb{R}$.
- More about how the computations are done, e.g. Mayer-Vietoris spectral sequence.
- More about barcodes: the stronger classification theorems, metrics on the space of barcodes, etc.
- Manifold estimation, dimension estimation.
- Creating smaller complexes for efficiency: Landmark points, Delaunay complex, α-complex, witness complex, etc.

Learning Seminar

We'll continue during the week of April 30 to May 4.

Who is interested in participating? Who volunteers to give a talk?

When should we meet? Riemann Surfaces is MWF 2:30-3:40 and Representation Theory is TuTh 10:00-11:45.