# Local-to-Global Methods for Topological Data Analysis

Brad Nelson

Institute for Computational and Mathematical Engineering
Stanford University

AI-at-SLAC Seminar
October 1, 2018

# Acknowledgments

The first half of this talk is intended to be an introduction to TDA, and is not original work (aside from defects). Several review papers now exist that cover the material from various perspectives including (but not limited to) [1, 2, 3].

Some of the examples in the first section are based on work with Rahul Sarkar (Stanford ICME)

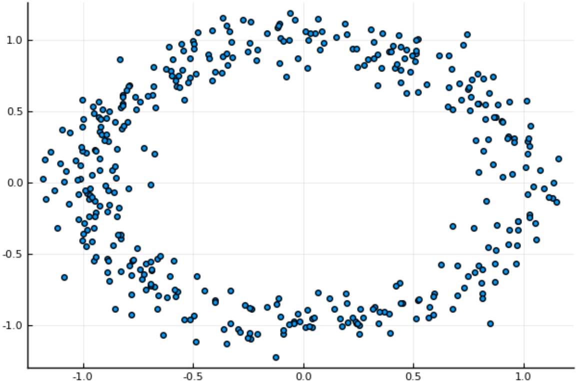
The second half of this talk is work with/supervised by my advisor, Gunnar Carlsson (Stanford Mathematics/Ayasdi)

# Outline

# Outline

“Data has shape, and shape has meaning”
Gunnar Carlsson

# What do you see?




Using Topological Data Analysis is one way to get a computer to agree with you.

# What is Topology?

The study of deformation invariants of spaces.
In topological data analysis, we often think

$$\text{Open Set} \leftrightarrow \text{Similarity Neighborhood}$$

A similarity can be as strong as a metric, or as weak as an equivalence relation.

# A Bird's Eye View of Algebraic Topology

- Study topological spaces using algebraic objects (groups, rings, modules, vector spaces)
  - Example: Fundamental Group
  - Example: Homology
- Primarily interested in *topological invariants*
  - Deforming or continuously modifying a space doesn't change result
- Interested in *functors*
  - Maps between spaces become maps between algebraic objects

$$\text{Algebra} = \text{Computable}$$

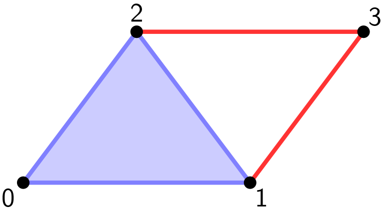
# Simplicial Complexes

**Question:** How to encode a topological space on a computer?

(One) **Answer:** A simplicial complex

# Simplicial Complexes - I

Key: Blue denotes Contractible, Red denotes a hole

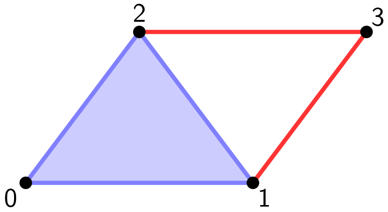



Simplicial complexes are one way to generalize graphs.
Let $X$ be a simplicial complex. Then $X$ has data:

- a vertex set $\{v_0, \ldots, v_n\}$
- a collection of simplices on the vertex set
  - a $k$-simplex $\sigma$ spans $k+1$ vertices and is denoted $\sigma = (v_{i_0}, \ldots, v_{i_k})$

**Example:** An undirected graph $G(V, E)$ (with no self edges) is a simplicial complex. $V$ is the vertex set, and $E$ is a set of $1$-simplices

# Simplicial Complexes - II



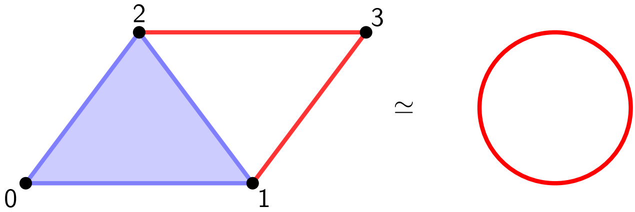

Let $X_k$ denote the set of $k$-simplices in $X$.

$$X_0 = \{(0), (1), (2), (3)\}$$
$$X_1 = \{(0,1), (0,2), (1,2), (1,3), (2,3)\}$$
$$X_2 = \{(0,1,2)\}$$

A **face** of a simplex $\sigma$ is a simplex in the boundary $\partial\sigma$ that is one dimension smaller (remove a single vertex from $\sigma$).

# Homology

Now we want a way to say something about the topology of the space.




One way to do this is to use **homology**.

# Homology - I

A **chain complex** is a sequence of vector spaces with boundary operators

$$\to C_{k+1} \xrightarrow{\partial_{k+1}} C_k \xrightarrow{\partial_k} C_{k-1} \to \dots$$

where $\partial_k \circ \partial_{k+1} = 0$.
One way to get a chain complex is to start with a simplicial complex $X$ and a field $F$. Recall

$$X_k = \{k\text{-simplices in } X\}$$

And let

$$C_k = \text{Free vector space over } X_k$$

(Free refers to a vector space with basis given by the elements of $X_k$).
The boundary operator $\partial_k$ will send a $k$-simplex to a $F$-linear combination of its faces.
**Example:** The incidence matrix of a graph $G$ is $\partial_1$

# Homology - II

Specifically, let $(v_0, \dots v_k)$ be a $k$-simplex, then $\alpha(v_0, \dots, v_k) \in C_k$, where $\alpha \in F$. Then

$$\partial_k \alpha(v_0, \dots, v_k) = \alpha \sum_{i=0}^{k} (-1)^i (v_0, \dots, \hat{v}_i, \dots, v_k)$$

where $\hat{v}_i$ indicates we have the face where $v_i$ is not present.
If $k = 0$, then $\partial_0(v_0) = 0$.

$$\partial(\bullet \!\longrightarrow\! \circ) = (\circ) - (\bullet)$$

One can verify that $\partial_k \partial_{k+1} = 0$. How?
To go from a $k$-simplex to a $(k-2)$-simplex, we remove two vertices. We can do this in two orders, each of which produces the opposite sign.

# Homology - III

The property $\partial_k \partial_{k+1} = 0$ means

$$\ker(\partial_k) \subseteq \operatorname{img}(\partial_{k+1})$$

In $C_k$, we call $\ker(\partial_k)$ "cycles", and $\operatorname{img}(\partial_{k+1})$ "boundaries". Note that these are both vector sub-spaces of $C_k$.

The perimeter of the triangle is always a cycle.
However, only on the left it is also a boundary.

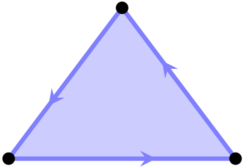
The picture suggests that cycles that aren't boundaries are topologically meaningful.

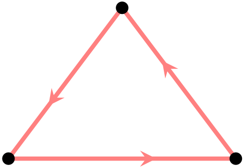
# Homology - IV

Mathematically, we have a way of describing cycles that aren't boundaries using a quotient vector space. Define

$$H_k(X) = \ker(\partial_k)/\operatorname{img}(\partial_{k+1})$$

the dimension of $H_k$ is also known as the $k$th betti number $\beta_k(X)$.





$$H_k(X) = \begin{cases} F & k = 0 \\ 0 & k = 1 \end{cases}$$

$$H_k(X) = \begin{cases} F & k = 0 \\ F & k = 1 \end{cases}$$

- $\dim(H_0(X))$ is number of connected components of $X$
- $\dim(H_1(X))$ is number of "loops" in $X$
- $\dim(H_k(X))$ is number of $k$-dimensional "voids" in $X$

# Persistent Homology - I

Often in applications, we don't like to choose parameters.

If there's a single parameter we vary, this gives rise to a **filtered space**

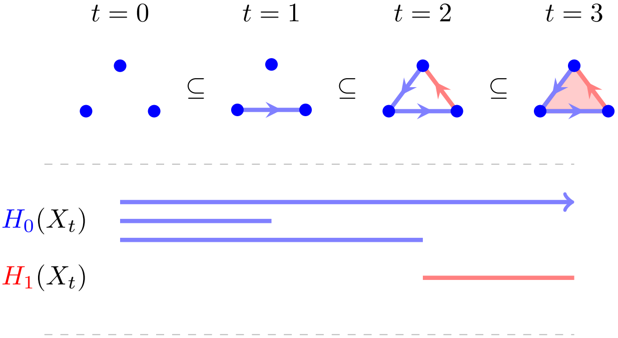
$$X_t = \cdots \subseteq X_{t-\epsilon} \subseteq X_t \subseteq X_{t+\epsilon} \subseteq \ldots$$

The parameter can be anything - height, curvature, distance - any real-valued function on data.
**Persistent Homology** $PH(X_t)$ tracks the *birth* and *death* of homology throughout this filtration, as the space changes.

This is one way topologists can tell coffee mugs from doughnuts.

# Persistent Homology - II

A filtration $X_t$ produces a *Persistence Barcode*.




| | $t=0$ | $t=1$ | $t=2$ | $t=3$ |
|---|---|---|---|---|
| $H_0(X_t)$ | $F^3$ | $F^2$ | $F$ | $F$ |
| $H_1(X_t)$ | $0$ | $0$ | $F$ | $0$ |

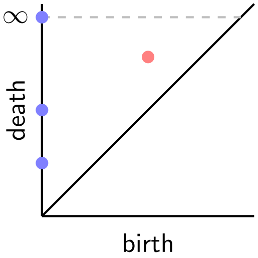
Top: Filtration. Middle: Barcode. Bottom: Homology vector spaces.

# Persistent Homology - III

The information in a barcode is a set of birth-death pairs.


$$\text{Pairs} = \left\{(0,1), (0,2), (0,\infty), (2,3)\right\}$$

We can also encode this information in a *Persistence Diagram*.

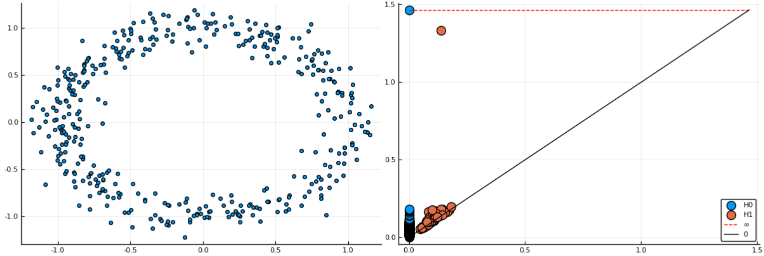


# Vietoris-Rips Filtrations - I

One easy filtration to define is the Vietoris-Rips Filtration.

1. Start with a finite dimensional vector space on $n$ points.
   This is the vertex set.
2. At time $r$, we put an edge between $v, w \in V$ if $d(v, w) \leq r$
3. At time $r$, we have a $k$-simplex $(v_0, \dots v_k)$, if the vertices form a $(k+1)$-clique.

# Vietoris-Rips Filtrations - II




Left: Point Cloud

Right: Persistence Diagram for Vietoris-Rips Filtration

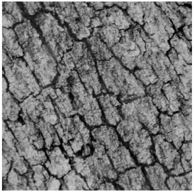
# Image-Based Filtrations - I

One way to define a filtration on a (gray-scale) image is the following:
Let $f(p)$ denote the (gray-scale) intensity at a pixel $p$.

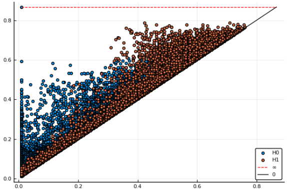
1. Make the vertex set the set of pixels in the image
2. A pixel $p$ is in the filtration at time $t$ if $f(p) \leq t$
3. Each pixel in the filtration connects to neighbors that are also in the filtration.

This is an example of a sub-levelset filtration.

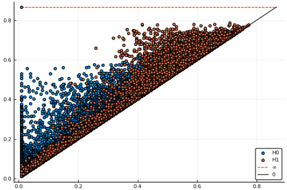
# Image-Based Filtrations - II



⇒




⇒




Persistence diagrams from Sipi texture rotation dataset.
Joint work with Rahul Sarkar.

# Featurization - I

Persistence diagrams/barcodes are not immediately useful for machine learning. We face the following challenges when comparing two diagrams:

- The number of birth-death pairs may differ between diagrams
- There's not a natural vector-space structure (can't throw into neural net, SVM, regression, etc.)

There have been a variety of solutions proposed:

- Polynomial featurization [4]
- Persistence Images [5]
- Persistence Landscapes [6]
- ...

# Featurization - II

Polynomial featurization[4]. For each homology dimension:

1. Transform pairs $(b_i, d_i) \mapsto (m_i, \ell_i)$, where

$$m_i = (b_i + d_i)/2$$
$$\ell_i = d_i - b_i$$

2. Evaluate a polynomial function of these pairs

$$p_\alpha(\{(m_i, \ell_i)\}) = \frac{1}{n}\Big[\sum_{k_1,k_2} \alpha_{k_1,k_2} \sum_i m_i^{k_1} \ell_i^{k_2}\Big]$$

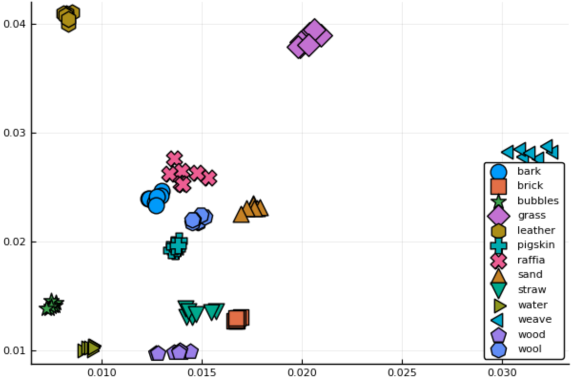
   where $n$ is the number of pairs (note these are like moments of persistence diagrams).

Do this for a few polynomials, and you now have several real numbers that are a function of a barcode.

- Same number of features, no matter how many birth-death pairs
- Naturally lives in $\mathbb{R}^k$ ($k$ is number of polynomials).

# Featurization - III

Two polynomial features on Sipi texture recognition data.




Horizontal - 1st mixed moment in $H_0$.
Vertical - 1st mixed moment in $H_1$.
Joint work with Rahul Sarkar.

# More Applications

Ingredients: Data + Filtration (+ Featurization + ML)

- Materials Discovery - Hiraoka [5]
  - Use of filtrations similar to the one presented for images
- Drug Discovery - G. Wei [7]
  - Molecules are treated as topological spaces
  - Filtrations similar to Rips construction
- Shape Classification [8]
  - Shapes treated as topological spaces
  - Filtrations based on coordinate axis, curvature, etc.

(This is not a complete list, even for applications described.)

# Outline

# What is Topological Modeling?

We're looking for a topological space that lies near sampled data.
Structures of interest:

- Parameterization of the space (strong result)
- Clusters, flares
- Maps to other spaces

**Examples:**

- Models of Image Patches (this talk)
- Dynamical systems [9]

# What is an Image Patch?

Image from van Hateren Natural Images Database [12]

- A $d \times d$ pixel block of a natural image
- This talk: black & white images
  - Pixel values measure light intensity
- We're interested in "high contrast" image patches

# Why Study Image Patches?

Natural images are complex. Patches are (relatively) simple.

A sparse, rough and incomplete picture:

- Image Compression, Harmonic Analysis
  - Ridgelets, Curvelets, etc. [10]
- Image denoising/interpolation [11]
- Vision - understand the visual cortex [12, 13]

# Image Patches in TDA

Circles and Klein Bottles

- Carlsson & de Silva (2004) use as an example in witness complex construction. [14]
  - 3 circle model for a dense region of $3 \times 3$ patches
- Carlsson et al. (2008) formulate Klein bottle model. [15]
  - Showed that a dense region of $3 \times 3$ patches lies near a Klein bottle
- Adams & Carlsson (2009) - range images primary circle [16]

Applications:

- Compression scheme by Maleki, Shahram, Carlsson (2008): BiWedgelets / Kleinlets [17]
- Perea & Carlsson (2014) - rotation-invariant pattern recognition using Klein bottle [18]

# What's New?

Contributions in this talk:

- Fiber bundle construction for the image patch Klein bottle
- generalization of this construction to higher dimensional data

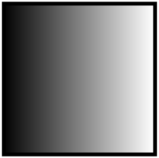
# Whence the Klein Bottle?

???

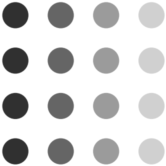
Klein Bottle artist: Ron Estrin

# Our Viewpoint

- Image patches are functions (intensity) on a $d \times d$ grid
- Image patches are a discretization of a continuous phenomenon



$f(\mathbf{x})$



$f(\mathbf{x_i})$

# Obtaining Patches

1. Randomly select 5000 patches from image
2. Take log intensities of patches
3. mean-center patches
4. take top 20% by contrast norm
5. normalize by contrast norm

Procedure from Lee, Pedersen, Mumford (2003) [13] applied to van Hateren dataset [12]

# Edges

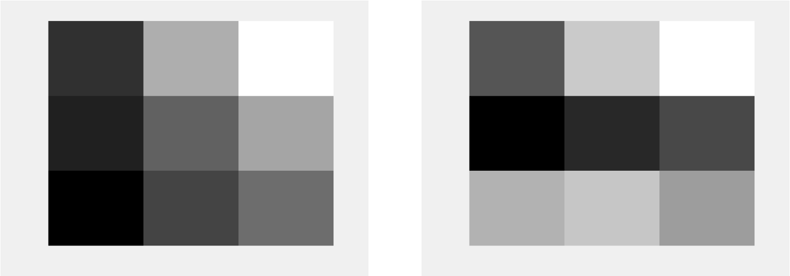
Our model for “high-contrast” image patches is based on the observation that patches that look like edges are common.

Example image patches from the van Hateren data set [12]

# Sources of Variation

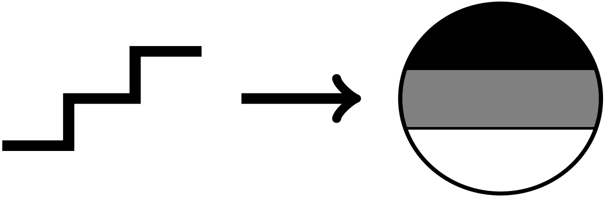
Two sources of variation in edges:

- Orientation of edges (primary circle)
- Type of edge (secondary circles)



Example image patches from the van Hateren data set [12]

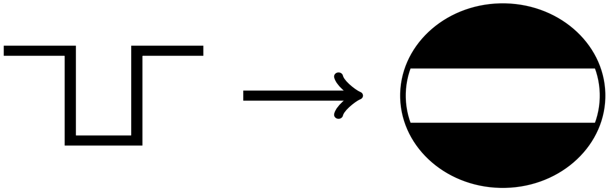
# Odd Functions



Odd functions capture the behavior of transitions from one intensity to another

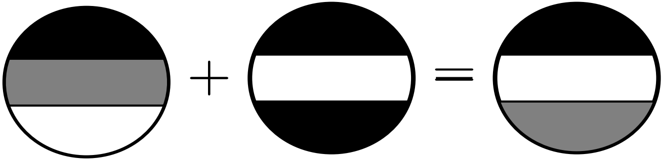
# Even Functions

Even functions capture the behavior of lines



# Combined Functions

Edges can also be a combination of even and odd functions



Generally, a sampled patch will have both an even and odd component.

# Density & Topology

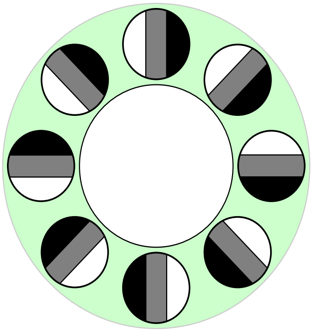
Even though we've already restricted ourselves to "high contrast" image patches, samples will eventually fill in the space.

We're interested in modeling "high-density" subspaces.

- Compute density estimator at each sample (using knn)
- Filter data to contain top $p\%$ by density
- compute persistent homology
- relax the density threshold and repeat

# Circle of Odd Functions

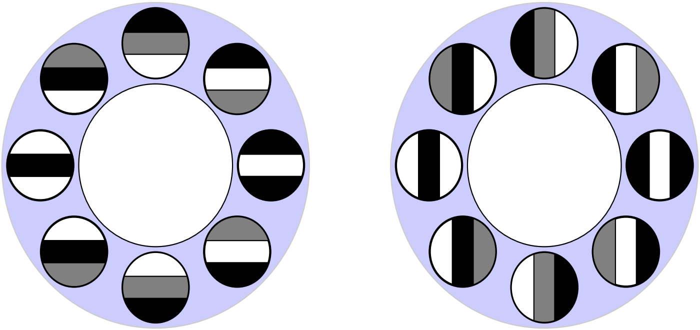
The first topological feature to be computed in the $3 \times 3$ image patch data set was a circle of odd functions [14]. This is the "most dense" portion of the data.



(This is the "Primary Circle" [14])

# Mixing Even and Odd

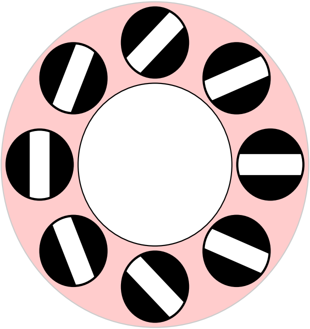
Other features that appear as the density threshold is relaxed are circles through even and odd functions at a fixed orientation.



(These are “Secondary Circles” [14])

# Circle of Even Functions

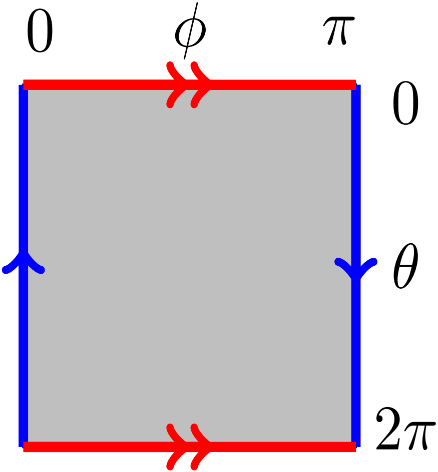
In larger patch sizes ($4 \times 4$ and larger), you can also find patches that live on a circle of even functions.



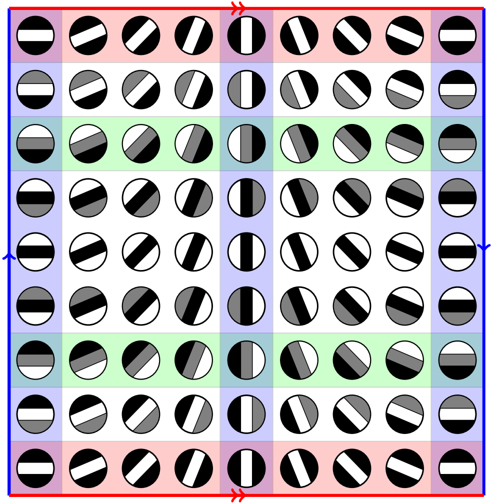
Note that one rotation in this circle is half a rotation in angle.

# Identification

The Klein bottle can be obtained by making an identification on the square.



# Klein Bottle Identification



Green: primary circle, Blue: secondary circles

# Outline

# How to Generalize?

Finding the Klein bottle in image patches took some time.

- How can this be turned into a general process?
- What were the important ingredients?

# How to Generalize?

Finding the Klein bottle in image patches took some time.

- How can this be turned into a general process?
- What were the important ingredients?

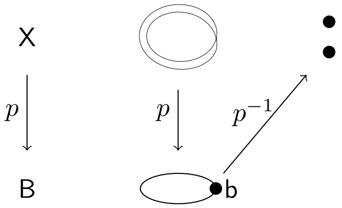
Some (possible) answers:

- Find easy (simple) structure first (circle of edges).
  - Build on this to add complexity
- One important ingredient was an orientation **map**

We're going to focus on leveraging **maps** to learn about spaces

# Maps and Fibers

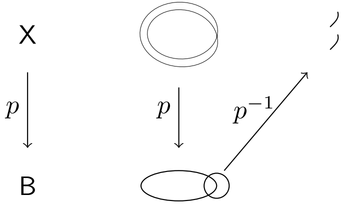
A **map** is a continuous function. Let $p : X \to B$ be a map.
A **fiber** of the map is the inverse image of a point $b \in B$, $p^{-1}(b) \subseteq X$.




Example: $X = S^1$, $B = S^1$, $p : \theta \mapsto 2\theta$
$p^{-1}(\theta) = \{\frac{\theta}{2}, \frac{\theta}{2} + \pi\}$

# Fiber Bundles

A **fiber bundle** is a map $p : X \to B$, where inverse images look locally like a product of an open set in $B$ with a space $F$.[1]




Example: $X = S^1$, $B = S^1$, $p : \theta \mapsto 2\theta$
$p^{-1}(\theta) = \{\frac{\theta}{2}, \frac{\theta}{2} + \pi\}$, $F$ is a two-point space

$F$ is called the fiber, $B$ is called the base space, and $X$ is called the total space.

[1]This isn't the full definition of a fiber bundle

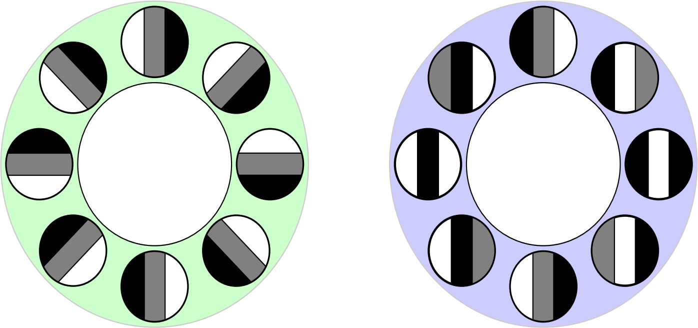
# Fiber Bundles - II

**Examples**:

- $B = S^1$, $F = [-1, 1]$
  - Cylinder (trivial)
  - Möbius Band (twisted)
- $B = S^1$, $F = S^1$
  - Torus (trivial)
  - Klein bottle (twisted)

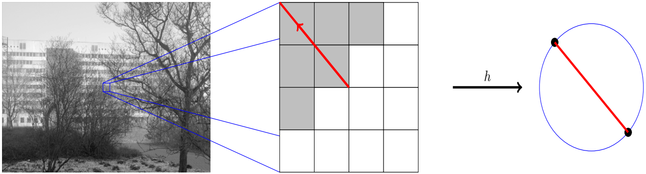
# Image Patches Revisited - I

A circle of edges at various orientations was the first discovery.



The second discoveries were circles of edges at a fixed orientation.
Idea: an **orientation map** can reveal a fiber bundle structure.

# Orientation Estimation - Harris Matrix [19]




$$I_1^2(x) = \frac{1}{k(k-1)} \sum_{j=1}^{k} \sum_{i=1}^{k-1} (x_{i,j} - x_{i+1,j})^2$$

$$I_2^2(p) = \frac{1}{k(k-1)} \sum_{i=1}^{k} \sum_{j=1}^{k-1} (x_{i,j} - x_{i,j+1})^2$$

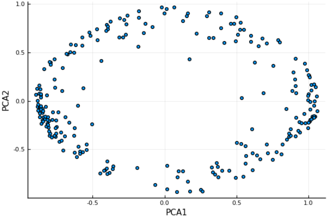
$$I_{12}(x) = \frac{1}{(k-1)^2} \sum_{i=1}^{k-1} \sum_{j=1}^{k-1} (x_{i,j} - x_{i,j+1})(x_{i,j} - x_{i+1,j})$$

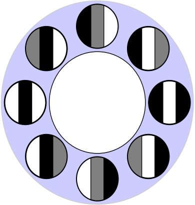
$$v_\phi(x) = \text{MaxEigVec} \begin{bmatrix} I_1^2 & I_{12} \\ I_{12} & I_2^2 \end{bmatrix}$$

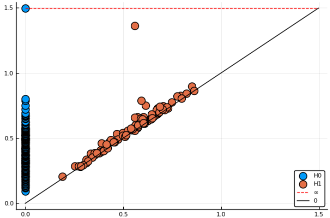
- Sign ambiguity in eigenvector naturally identifies antipodal points on circle: $v_\phi \sim -v_\phi$
- $(S^1/\sim) \simeq \mathbb{R}P^1 \simeq S^1$
- map is similar to: $\phi \mapsto 2\phi$

# Image Patches Revisited - II

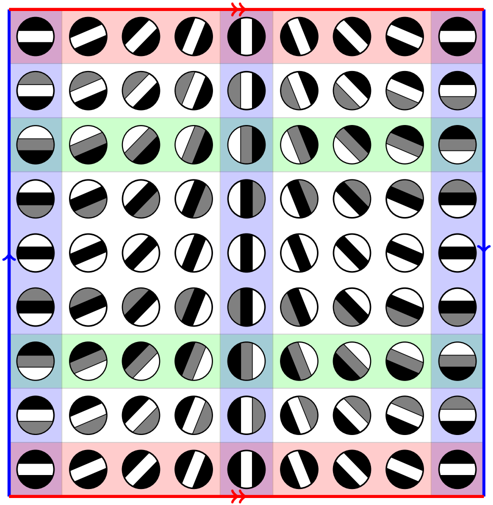
Point cloud is a fiber of the Harris Map ($7 \times 7$ patches). Inverse image of an open set covering about 10% of $\mathbb{R}P^1$





# Image Patches Revisited - III

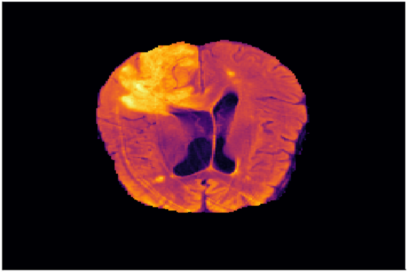


Vertical - Fibers. Horizontal - Section of Base Space

# 3D Image Patches

A similar, but more complex, set of data comes from 3-dimensional images.

- pixels $\Rightarrow$ voxels
- patches $\Rightarrow$ $d \times d \times d$ blocks of voxels



One source of 3D-images: MRI [20]

# 3D Orientations

We used an open-source MRI dataset (OASIS)[20] to use for sampling

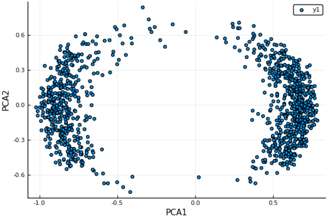
- Initial investigations revealed a sphere of odd edges.
- We then decided to investigate fibers of the orientation map.
- Note that in 3 dimensions, the space of orientations is

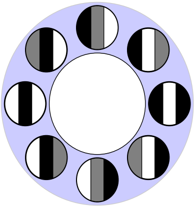
$$(S^2/\sim) = \mathbb{R}P^2$$

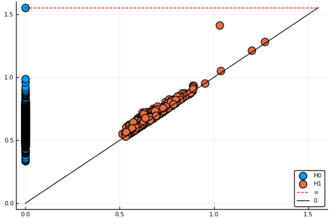
- Harris map generalizes to higher dimensions
  - in $d$ dimensions - eigenvector of $d \times d$ matrix
  - finite differences in each dimension

# 3D Orientation Fibers

Point cloud is a fiber of the Harris Map ($7 \times 7 \times 7$ cubes).
Inverse image of an open set in $\mathbb{R}P^2$.





# Stitching Everything Together

We've seen some evidence that fibers of the orientation map again look like circles.

What is the topological model for this?

- A fiber bundle with fiber $S^1$ and base space $\mathbb{R}P^2$
- As in the Klein bottle, there are twists
- The Klein bottle is a subspace (diameters)

**Ongoing work:** We're investigating ways to use structure to speed up computations.

# Conclusion

- Overview of Persistent Homology, some applications
- Topological modeling of data
  - Klein bottle in image patches
- Using maps and fibers to investigate spaces
  - Proposed model for 3D image patches

**What's Next?**

- (ongoing) Computational framework that exploits structure of maps
- Applications that use structure of 3D patch space
  - Can we compress 3D images?
  - Can we understand 3D CNNs?
- Interested in finding other examples of fiber bundles in data

# References I


- G. Carlsson. Topology and Data. Bulletin of the AMS. 46(2). 2009.
- R. Ghrist. Homological Algebra and Data. Preprint: https://www.math.upenn.edu/ ghrist/preprints/HAD.pdf 2017.
- N. Otter et. al. A roadmap for the computation of persistent homology. EPJ Data Science 6(1). 2017.
- A. Adcock, E. Carlsson, G. Carlsson. The ring of algebraic functions on persistence bar codes. Homology, Homotopy and Applications. 18(1) 2016.
- I. Obayashi and Y. Hiraoka. Persistence Diagrams with Linear Machine Learning Models. arXiv:1706.10082 2017.
- P. Bubenik. Statistical Topological Data Analysis using Persistence Landscapes. JMLR 16 pp. 77-102. 2015.
- Z. Cang and G. Wei. TopologyNet: Topology based deep convolutional and multi-task neural networks for biomolecular property predictions. PLOS Computational Biology. 13(7). 2017.

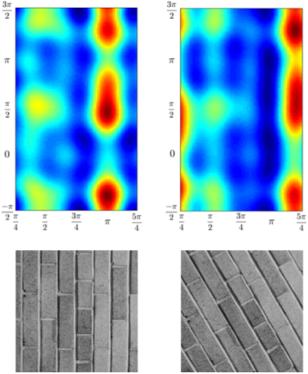
# References II


- F. Chazal et. al. Gromov-Hausdorff Stable Signatures for Shapes using Persistence. Computer Graphics Forum 28. pp. 13931403. 2009.
- J. Perea and J. Harer. Sliding Windows and Persistence: An Application of Topological Methods to Signal Analysis. FOCM 15(3). pp. 799-838. 2015.
- D.L. Donoho, A.G. Flesia, Can recent innovations in harmonic analysis explain key findings in natural image statistics?. Network: computation in neural systems, 12(3), pp.371-393. 2001.
- S. Osher, Z. Shi, and W. Zhu. Low Dimensional Manifold Model for Image Processing. Technical Report UCLA CAM Report 16-04, 2016.
- J.H. van Hateren and A. van der Schaaf. Independent component filters of natural images compared with simple cells in primary visual cortex. Proc. R. Soc. Lond., B265:359366, 1998.
- A.B. Lee, K.S. Pedersen, and D.B. Mumford. The nonlinear statistics of high-contrast patches in natural images. International Journal of Computer Vision, 54:83103, 2003.

# References III


- V. de Silva, G. Carlsson. Topological estimation using witness complexes. Eurographics Symposium on Point-Based Graphics. 2004.
- G. Carlsson, T. Ishkhanov, V. de Silva, and A. Zomorodian. On the local behavior of spaces of natural images. International Journal of Computer Vision, 76(1):112, 2008.
- H. Adams, G. Carlsson. On the nonlinear statistics of range image patches. SIAM J. Imaging Sciences, 2(1):110117, 2009.
- A. Maleki, M. Shahram, G. Carlsson A Near Optimal Coder For Image Geometry With Adaptive Partitioning. 15th IEEE International Conference on Image Processing, 2008.
- J. Perea, G. Carlsson. A klein-bottle-based dictionary for texture representation. International Journal of Computer Vision, 107(1):7597, 2014.
- C. Harris, M. Stephens. A combined corner and edge detector. 1988.
- Marcus, D. S. et al. (2010). Open Access Series of Imaging Studies (OASIS): Longitudinal MRI Data in Nondemented and Demented Older Adults. Journal of cognitive neuroscience 22.12: 26772684.

# Texture Recognition



J. Perea, G. Carlsson "A Klein-Bottle-Based Dictionary for Texture Representation" (2014)

# Compression


A. Maleki, M. Shahram, G. Carlsson "A Near Optimal Coder For Image Geometry With Adaptive Partitioning" (2008)