The Shape of Biomedical Data
ACM-BCB 2016

Gunnar Carlsson
Stanford University and Ayasdi Inc.

October 2, 2016
Big Data

It's not all about the "Big"
Big Data

- Complexity is a fundamental issue
Big Data

- Complexity is a fundamental issue
- Complexity both in structure and format
Big Data

- Complexity is a fundamental issue
- Complexity both in structure and format
- Requires an organizing principle
Shape of Data

- Data has shape
Shape of Data

- Data has shape
- The shape matters
Shape of Data

Linear Regression
Shape of Data

Clusters
Shape of Data

Loop
Shape of Data

“Y-junction”
Shape of Data

- How to model data?
How to model data?
Usually done algebraically - lines, quadratics, etc.
Shape of Data

- How to model data?
- Usually done algebraically - lines, quadratics, etc.
- Capturing all kinds of shape requires different method
Shape of Data

- How to model data?
- Usually done algebraically - lines, quadratics, etc.
- Capturing all kinds of shape requires different method
- Topological modeling
Shape of Data

- Normally defined in terms of a distance metric
Shape of Data

- Normally defined in terms of a distance metric
- Euclidean distance, Hamming, correlation distance, etc.
Shape of Data

- Normally defined in terms of a distance metric
- Euclidean distance, Hamming, correlation distance, etc.
- Encodes similarity
Topology

- Formalism for measuring and representing shape
Topology

- Formalism for measuring and representing shape
- Pure mathematics since 1700’s
Topology

- Formalism for measuring and representing shape
- Pure mathematics since 1700’s
- Last ten years ported into the point cloud world
Topology

Königsberg Bridges
Topology

Three key ideas:
Topology

Three key ideas:

- Coordinate freeness
Topology

Three key ideas:

▶ Coordinate freeness
▶ Invariance under deformation
Topology

Three key ideas:

- Coordinate freeness
- Invariance under deformation
- Compressed representations
Topology

Coordinate Freeness
Topology

Invariance to Deformations
Topology

Coffee cup is the “same” as a doughnut
Topology

Coffee cup is the “same” as a doughnut
Coffee cup is the “same” as a doughnut
Topology

Coffee cup is the "same" as a doughnut
Coffee cup is the “same” as a doughnut
Coffee cup is the “same” as a doughnut
Topology

Coffee cup is the “same” as a doughnut
Topology

Coffee cup is the “same” as a doughnut
Coffee cup is the “same” as a doughnut
Coffee cup is the “same” as a doughnut
Coffee cup is the “same” as a doughnut
Topology

Coffee cup is the “same” as a doughnut
Coffee cup is the “same” as a doughnut
Topology

Coffee cup is the “same” as a doughnut
Topology

Coffee cup is the “same” as a doughnut
Coffee cup is the “same” as a doughnut
Topology

Coffee cup is the “same” as a doughnut
Coffee cup is the “same” as a doughnut
Topology

Coffee cup is the “same” as a doughnut
Coffee cup is the “same” as a doughnut
Coffee cup is the “same” as a doughnut
Coffee cup is the “same” as a doughnut
Coffee cup is the “same” as a doughnut
Topology

Coffee cup is the “same” as a doughnut
Topology

Coffee cup is the “same” as a doughnut
Topology

Coffee cup is the “same” as a doughnut
Topology

Coffee cup is the “same” as a doughnut
Coffee cup is the “same” as a doughnut
Coffee cup is the “same” as a doughnut
Coffee cup is the “same” as a doughnut
Coffee cup is the “same” as a doughnut
Topology

Coffee cup is the “same” as a doughnut
Coffee cup is the “same” as a doughnut
Coffee cup is the “same” as a doughnut
Topology

Coffee cup is the “same” as a doughnut
Coffee cup is the “same” as a doughnut
Coffee cup is the “same” as a doughnut
Topology

Coffee cup is the “same” as a doughnut
Coffee cup is the “same” as a doughnut
Coffee cup is the “same” as a doughnut
Topology

Coffee cup is the “same” as a doughnut
Coffee cup is the “same” as a doughnut
Topology

Compressed Representations of Geometry
Topology

Two tasks:
Topology

Two tasks:

- Represent shape
Topology

Two tasks:

- Represent shape
- Measure shape
Can one extend topological mapping methods (compressed representations) from idealized shapes to data?
Can one extend topological mapping methods (compressed representations) from idealized shapes to data?

Yes (Singh, Memoli, G. C.)
Topological Mapping

Covering of Circle
Topological Mapping

Create nodes
Topological Mapping

Create edges
Topological Mapping

Nerve complex
Mapping

Now given point cloud data set $X$, and a covering $U$. 
Mapping

Now given point cloud data set $\mathbb{X}$, and a covering $\mathcal{U}$. 

Build simplicial complex same way, but components replaced by clusters.
How to choose coverings?
How to choose coverings?

Given a reference map (or filter) $f : X \to Z$, where $Z$ is a metric space, and a covering $\mathcal{U}$ of $Z$, can consider the covering $\{f^{-1}U_\alpha\}_{\alpha \in A}$ of $X$. Typical choices of $Z$ - $\mathbb{R}$, $\mathbb{R}^2$, $S^1$. 
How to choose coverings?

Given a reference map (or filter) $f : \mathbb{X} \to Z$, where $Z$ is a metric space, and a covering $\mathcal{U}$ of $Z$, can consider the covering $\{f^{-1}U_\alpha\}_{\alpha \in A}$ of $\mathbb{X}$. Typical choices of $Z$ - $\mathbb{R}$, $\mathbb{R}^2$, $S^1$.

The reference space typically has useful families of coverings attached to it.
Mapping
Typical one dimensional filters:

- Density estimators
Typical one dimensional filters:

- Density estimators
- Measures of data depth, e.g. \( \sum_{x' \in X} (d(x, x'))^2 \)
Typical one dimensional filters:

- Density estimators
- Measures of data depth, e.g. \( \sum_{x' \in X} d(x, x')^2 \)
- Eigenfunctions of graph Laplacian for Vietoris-Rips graph
Mapping

Typical one dimensional filters:

- Density estimators
- Measures of data depth, e.g. $\sum_{x' \in X} d(x, x')^2$
- Eigenfunctions of graph Laplacian for Vietoris-Rips graph
- PCA or MDS coordinates
Mapping

Typical one dimensional filters:

- Density estimators
- Measures of data depth, e.g. $\sum_{x' \in X} d(x, x')^2$
- Eigenfunctions of graph Laplacian for Vietoris-Rips graph
- PCA or MDS coordinates
- User defined, data dependent filter functions
Mapping

Relationships between diabetic, pre-diabetic and healthy populations

Glucose level

Insulin response

Healthy

Pre-diabetic

Overt Diabetic

Miller-Reaven Diabetes Dataset
Mapping

Li et al, Science Translational Medicine, 2015
Mapping

Torres et al, PLOS Biology, 2016
Mapping

Uninfected
No treatment
Ampicillin treatment
Hsp 26
AttA
CFU
CG3117 CG32444
A B
C
D E

i.
ii.
iii.
iv.
v.

Uninfected
Acute response
Sickness
Death
Recovery

Figure 4
1. Heat shock proteins
2. Secretory apparatus
3. Antimicrobial peptides
4. Metabolic changes
5. Mitochondrial respiration
6. Proteasome

Low
High

Louie et al, PLOS Biology, 2016
Economic Regime Analysis
Mapping

Economic Regime Analysis
Mapping

RNA hairpin folding data
Diagram of gene expression profiles for breast cancer
M. Nicolau, A. Levine, and G. Carlsson, PNAS 2011
Mapping

Comparison with hierarchical clustering
Different platforms - importance of coordinate free approach
Mapping

Serendipity - copy number variation reveals parent child relations
Example: Quality Control

Data handling is not an error-free process; mislabeling control samples can lead to incorrect assumptions in your analysis. Within minutes, Ayasdi Iris identified a sub-structure separated from the rest of the network. Initially thought to be a specific treatment with stark differences in cell effects, a deeper look at the well locations showed that these were mislabeled control samples.

About the Data
In an experiment testing cell exposure to various bacterial strains, cells were separated in a 96-well plate, the labeling of which was done by hand in the lab.

8,022 cell samples
8187 measurements
Topological Modeling

- Suggests a new kind of modeling
Topological Modeling

- Suggests a new kind of modeling
- Output is no longer a set of algebraic formulae, but a network
Topological Modeling

- Suggests a new kind of modeling
- Output is no longer a set of algebraic formulae, but a network
- Input is a finite set equipped with a distance function
Topological Modeling

- Suggests a new kind of modeling
- Output is no longer a set of algebraic formulae, but a network
- Input is a finite set equipped with a distance function
- Distance function encodes similarity
Topological Modeling

\[ y = 1 - x \]
### Topological Modeling

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 1</td>
<td>100</td>
<td>5.123</td>
<td>17.89</td>
<td>0.05589755</td>
</tr>
<tr>
<td></td>
<td>101</td>
<td>199.75</td>
<td>36.52</td>
<td>0.02738226</td>
</tr>
<tr>
<td></td>
<td>102</td>
<td>201.75</td>
<td>73.78</td>
<td>0.01355381</td>
</tr>
<tr>
<td></td>
<td>103</td>
<td>203.75</td>
<td>148.3</td>
<td>0.00674302</td>
</tr>
<tr>
<td></td>
<td>104</td>
<td>205.75</td>
<td>297.34</td>
<td>0.00336315</td>
</tr>
<tr>
<td></td>
<td>105</td>
<td>207.75</td>
<td>595.42</td>
<td>0.00167949</td>
</tr>
<tr>
<td></td>
<td>106</td>
<td>209.75</td>
<td>1191.58</td>
<td>0.00082922</td>
</tr>
<tr>
<td></td>
<td>107</td>
<td>211.75</td>
<td>2383.9</td>
<td>0.000412958</td>
</tr>
<tr>
<td></td>
<td>108</td>
<td>213.75</td>
<td>4768.56</td>
<td>0.00020971</td>
</tr>
<tr>
<td></td>
<td>109</td>
<td>215.75</td>
<td>9537.82</td>
<td>0.00010685</td>
</tr>
<tr>
<td></td>
<td>205</td>
<td>217.75</td>
<td>1907.38</td>
<td>0.00457143</td>
</tr>
<tr>
<td></td>
<td>208</td>
<td>409.75</td>
<td>38153.5</td>
<td>0.002434527</td>
</tr>
<tr>
<td></td>
<td>211</td>
<td>415.75</td>
<td>76307.76</td>
<td>0.00336952</td>
</tr>
<tr>
<td></td>
<td>214</td>
<td>421.75</td>
<td>152611.22</td>
<td>0.00336546</td>
</tr>
<tr>
<td></td>
<td>217</td>
<td>427.75</td>
<td>305233.18</td>
<td>0.00233236</td>
</tr>
<tr>
<td></td>
<td>220</td>
<td>433.75</td>
<td>610467.1</td>
<td>0.002230017</td>
</tr>
<tr>
<td></td>
<td>223</td>
<td>439.75</td>
<td>1220934.94</td>
<td>0.00226086</td>
</tr>
<tr>
<td></td>
<td>226</td>
<td>445.75</td>
<td>2251870.02</td>
<td>0.00223839</td>
</tr>
<tr>
<td></td>
<td>229</td>
<td>451.75</td>
<td>4883741.98</td>
<td>0.00220872</td>
</tr>
<tr>
<td></td>
<td>267</td>
<td>457.75</td>
<td>9767484.7</td>
<td>0.00217964</td>
</tr>
<tr>
<td></td>
<td>289</td>
<td>533.75</td>
<td>19534970.1</td>
<td>0.00187003</td>
</tr>
<tr>
<td></td>
<td>271</td>
<td>537.75</td>
<td>30066941</td>
<td>0.00186415</td>
</tr>
<tr>
<td></td>
<td>273</td>
<td>541.75</td>
<td>7819388.82</td>
<td>0.00184247</td>
</tr>
<tr>
<td></td>
<td>275</td>
<td>545.75</td>
<td>156279766</td>
<td>0.00182859</td>
</tr>
<tr>
<td></td>
<td>277</td>
<td>549.75</td>
<td>312559933</td>
<td>0.00181571</td>
</tr>
<tr>
<td></td>
<td>279</td>
<td>553.75</td>
<td>625115067</td>
<td>0.00180261</td>
</tr>
<tr>
<td></td>
<td>281</td>
<td>557.75</td>
<td>1250238135</td>
<td>0.00178971</td>
</tr>
<tr>
<td></td>
<td>283</td>
<td>561.75</td>
<td>2500476727</td>
<td>0.00177699</td>
</tr>
<tr>
<td></td>
<td>285</td>
<td>565.75</td>
<td>5000952545</td>
<td>0.00176445</td>
</tr>
<tr>
<td></td>
<td>287</td>
<td>569.75</td>
<td>1.00026+10</td>
<td>0.00175208</td>
</tr>
</tbody>
</table>
Topological Modeling
Topological Modeling

?
Topological Modeling - Coloring by Function Values

World Values Survey - 2000 U.S. Respondents
11 Questions on Trust in Institutions
Color by response to left/right preference
Topological Modeling - Coloring by Function Values

Coloring by sum of trust in all 11 institutions
Topological Modeling - Coloring by Function Values

Color by response to “Do you feel you have control over your life?”
Topological Modeling - Coloring by Function Values

Response to “Should employers favor native born employees in difficult economic times?”
Response to “How much faith do you have in the U.N.?”
Topological Modeling - Hot Spot Analysis

- Suppose we are given outcome of interest, such as “survival”, “revenue”, “fraud”, “Democrat/Republican”, etc.
Topological Modeling - Hot Spot Analysis

- Suppose we are given outcome of interest, such as “survival”, “revenue”, “fraud”, “Democrat/Republican”, etc.
- Coloring by average value of outcome on data points in node is useful.
Topological Modeling - Hot Spot Analysis

- Suppose we are given outcome of interest, such as “survival”, “revenue”, “fraud”, “Democrat/Republican”, etc.
- Coloring by average value of outcome on data points in node is useful
- Frequently discover “hot spots” of concentration of high values of the outcome
Suppose we are given outcome of interest, such as “survival”, “revenue”, “fraud”, “Democrat/Republican”, etc.

Coloring by average value of outcome on data points in node is useful.

Frequently discover “hot spots” of concentration of high values of the outcome.

Extremely useful information.
Example: Model Verification

Network of patients colored by the predicted survival (upper left, blue indicates good predicted survival) and actual survival (lower right, blue indicates good survival) – a group of patients was identified with good predicted survival but bad outcomes. Further analysis showed that missing data was misleading the model used to make survival predictions.

About the Data

When patients come to an emergent care facility, doctors need to assess priority and predict probability of survival with medical intervention.

Patient is quickly assessed for information about their condition: temperature, blood pressure, yes/no questions.
Topological Modeling - Hot Spot Analysis

Program Downgrades
Topological Modeling - Hot Spot Analysis

Program Downgrades
Topological Modeling - Hot Spot Analysis

But downgrades are different

Credit Risk Analysis

Upgrades are the same
Topological Modeling - Feature Selection

- It is often useful to consider a topological model of the space of columns rather than rows in a data set.
Topological Modeling - Feature Selection

- It is often useful to consider a topological model of the space of columns rather than rows in a data set.
- Density is an interesting feature in this space - one often needs to compensate for overrepresented features.
Topological Modeling - Feature Selection

- It is often useful to consider a topological model of the space of columns rather than rows in a data set.
- Density is an interesting feature in this space - one often needs to compensate for overrepresented features.
- Centrality also interesting - least central features may be of most interest.
It is often useful to consider a topological model of the space of columns rather than rows in a data set.

Density is an interesting feature in this space - one often needs to compensate for overrepresented features.

Centrality also interesting - least central features may be of most interest.

Hot spot analysis in columns is also useful.
Topological Modeling - Feature Selection

CCAR Stress Test Analysis Model
Measuring Shape

- Shape is nebulous concept
Measuring Shape

- Shape is nebulous concept
- Nevertheless very important to make precise
Measuring Shape

- Shape is nebulous concept
- Nevertheless very important to make precise
- Important to be able to “measure” it precisely in an appropriate sense
Measuring Shape

- Shape is nebulous concept
- Nevertheless very important to make precise
- Important to be able to “measure” it precisely in an appropriate sense
- Achieve by counting occurrences of patterns in an appropriate sense
Measuring Shape
Measuring Shape

Capturing obstacle by “lassoing”
Capturing obstacle by “lassoing”
Two different lassos capture same obstacle
Measuring Shape

Solve by introducing homotopy relation
Measuring Shape

Second different lasso
Measuring Shape

Adding two lassos together
Measuring Shape

Multiplying a lasso by 2
Measuring Shape

- Algebraic topology performs counts of occurrences of *equivalence classes of geometric patterns*
Measuring Shape

- Algebraic topology performs counts of occurrences of *equivalence classes of geometric patterns*
- Naive counting typically give infinite answers
Measuring Shape

- Algebraic topology performs counts of occurrences of equivalence classes of geometric patterns
- Naive counting typically give infinite answers
- Counting is done by computing dimensions of algebraic objects
Measuring Shape

\[ b_1 = 1 \quad b_1 = 0 \quad b_1 = 2 \]
\[ b_2 = 0 \quad b_2 = 1 \quad b_2 = 1 \]

\( b_i \) is the “\( i \)-th Betti number”
Measuring Shape

Counts the number of “$i$-dimensional holes”
Betti numbers are computed as dimensions of Boolean vector spaces (E. Noether)
Measuring Shape

- Betti numbers are computed as dimensions of Boolean vector spaces (E. Noether)
- \( b_i(X) = \dim H_i(X) \)
Measuring Shape

- Betti numbers are computed as dimensions of Boolean vector spaces (E. Noether)
- \( b_i(X) = \dim H_i(X) \)
- \( H_i(X) \) is *functorial*, i.e. continuous map \( f : X \to Y \) induces linear transformation \( H_i(f) : H_i(X) \to H_i(Y) \)
Measuring Shape

- Betti numbers are computed as dimensions of Boolean vector spaces (E. Noether)
- \( b_i(X) = \dim H_i(X) \)
- \( H_i(X) \) is functorial, i.e. continuous map \( f : X \to Y \) induces linear transformation \( H_i(f) : H_i(X) \to H_i(Y) \)
- Computation is simple linear algebra over fields or integers
Measuring Shape of Data

- Need to extend homology to more general setting including point clouds
Measuring Shape of Data

- Need to extend homology to more general setting including point clouds
- Method called *persistent homology*
Measuring Shape of Data

- Need to extend homology to more general setting including point clouds
- Method called *persistent homology*
- Developed by Edelsbrunner, Letscher, and Zomorodian and Zomorodian-Carlsson
Measuring Shape of Data

- How to define homology to point clouds sensibly?
Measuring Shape of Data

- How to define homology to point clouds sensibly?
- Finite sets are discrete
Measuring Shape of Data

- How to define homology to point clouds sensibly?
- Finite sets are discrete
- Statisticians knew what to do
Measuring Shape of Data

Dendrogram
Measuring Shape of Data

- Points are connected when they are within a threshold $\epsilon$
Measuring Shape of Data

- Points are connected when they are within a threshold $\epsilon$.
- Dendrogram gives a profile of the clustering at all $\epsilon$’s simultaneously.
Measuring Shape of Data

- Points are connected when they are within a threshold $\epsilon$
- Dendrogram gives a profile of the clustering at all $\epsilon$’s simultaneously
- Doesn’t require choosing a threshold
Measuring Shape of Data

- How to build spaces from finite metric spaces
Measuring Shape of Data

- How to build spaces from finite metric spaces
- Use the nerve of the covering by balls of a given radius $\epsilon$
Measuring Shape of Data
Measuring Shape of Data
Measuring Shape of Data

- Provides an increasing sequence of simplicial complexes
Measuring Shape of Data

- Provides an increasing sequence of simplicial complexes
- Apply $H_i$
Measuring Shape of Data

- Provides an increasing sequence of simplicial complexes
- Apply $H_i$
- Gives a diagram of vector spaces (Noether’s functoriality)

\[ V_0 \to V_1 \to V_2 \to V_3 \to \cdots \]
Measuring Shape of Data

- Provides an increasing sequence of simplicial complexes
- Apply $H_i$
- Gives a diagram of vector spaces (Noether’s functoriality)

$$V_0 \rightarrow V_1 \rightarrow V_2 \rightarrow V_3 \rightarrow \cdots$$

- Call such algebraic structures *persistence vector spaces*
Can we classify persistence vector spaces, up to isomorphism?

- Yes, analogous to classification of ordinary vector spaces by dimension.
- Classification parametrized by bar codes, i.e. finite collections of intervals.
- Readily computable due to the judicious use of higher algebra.
Measuring the Shape of Data

- Can we classify persistence vector spaces, up to isomorphism?
- Yes, analogous to classification of ordinary vector spaces by dimension
Measuring the Shape of Data

- Can we classify persistence vector spaces, up to isomorphism?
- Yes, analogous to classification of ordinary vector spaces by dimension
- Classification parametrized by *bar codes*, i.e. finite collections of intervals
Measuring the Shape of Data

- Can we classify persistence vector spaces, up to isomorphism?
- Yes, analogous to classification of ordinary vector spaces by dimension
- Classification parametrized by *bar codes*, i.e. finite collections of intervals
- Readily computable due to the judicious use of higher algebra
Measuring the Shape of Data - Barcodes
Measuring the Shape of Data - Barcodes

One dimensional barcode:
Measuring the Shape of Data - Barcodes
Measuring the Shape of Data - Barcodes

$\beta_1=3$
Measuring the Shape of Data - Barcodes
Measuring the Shape of Data - Barcodes

$\beta_1 = 2$
Application to Natural Image Statistics

With V. de Silva, T. Ishkanov, A. Zomorodian
An image taken by black and white digital camera can be viewed as a vector, with one coordinate for each pixel.
Natural Images

An image taken by black and white digital camera can be viewed as a vector, with one coordinate for each pixel.

Each pixel has a "gray scale" value, can be thought of as a real number (in reality, takes one of 255 values).
Natural Images

An image taken by black and white digital camera can be viewed as a vector, with one coordinate for each pixel.

Each pixel has a “gray scale” value, can be thought of as a real number (in reality, takes one of 255 values).

Typical camera uses tens of thousands of pixels, so images lie in a very high dimensional space, call it $\mathcal{P}$. 

Natural Images

D. Mumford: What can be said about the set of images $\mathcal{I} \subseteq \mathcal{P}$ one obtains when one takes many images with a digital camera?
Solution (Lee, Mumford, Pedersen): Study local structure of images statistically, where there is less variation.
Solution (Lee, Mumford, Pedersen): Study local structure of images statistically, where there is less variation.

Specifically, study $3 \times 3$ patches in the image.
Natural Images

**Solution (Lee, Mumford, Pedersen):** Study *local* structure of images statistically, where there is less variation

Specifically, study $3 \times 3$ patches in the image.

Study high *density* high *contrast* patches
Primary Circle

5 × 10^4 points, k = 300, T = 25

One-dimensional barcode, suggests $\beta_1 = 1$
Primary Circle

\[5 \times 10^4 \text{ points, } k = 300, \ T = 25\]

One-dimensional barcode, suggests \(\beta_1 = 1\)

Is the set clustered around a circle?
Primary Circle

PRIMARY CIRCLE
Three Circle Model

$5 \times 10^4$ points, $k = 15$, $T = 25$

One-dimensional barcode, suggests $\beta_1 = 5$
Three Circle Model

$5 \times 10^4$ points, $k = 15$, $T = 25$

One-dimensional barcode, suggests $\beta_1 = 5$

What's the explanation for this?
Three Circle Model
Three Circle Model

THREE CIRCLE MODEL
Red and green circles do not touch, each touches black circle
Three Circle Model
Three Circle Model

\[ \beta_1 = 5 \]
Does the data fit with this model?
Three Circle Model

SECONDARY CIRCLE
Three Circle Model
IS THERE A TWO DIMENSIONAL SURFACE IN WHICH THIS PICTURE FITS?
Klein Bottle

$4.5 \times 10^6$ points, $k = 100$, $T = 10$
Klein Bottle

\( \mathcal{K} \)- KLEIN BOTTLE
Klein Bottle

<table>
<thead>
<tr>
<th>$i$</th>
<th>0</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_i(\mathcal{K})$</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>
Klein Bottle

<table>
<thead>
<tr>
<th>$i$</th>
<th>0</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_i(\mathcal{K})$</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

Agrees with the Betti numbers we found from data
Klein Bottle

Identification Space Model
Klein Bottle

Identification Space Model
Klein Bottle

Do the three circles fit naturally inside $\mathcal{K}$?
Klein Bottle

PRIMARY CIRCLE

P

Q

PQ

Q

PRIMAY CIRCLE
Klein Bottle

SECONDARY
CIRCLES

R
R
S
S

SECONDARY
CIRCLES
Mapping Patches
Mapping Patches
Mapping Patches
Klein bottle makes sense in quadratic polynomials in two variables, as polynomials which can be written as

\[ f = q(\lambda(x)) \]

where

1. \( q \) is single variable quadratic
2. \( \lambda \) is a linear functional
3. \( \int_D f = 0 \)
4. \( \int_D f^2 = 1 \)
This understanding of density can be applied to develop compression schemes.
This understanding of density can be applied to develop compression schemes.

Earlier work, based on primary circle, called “Wedgelets”,
done by Baraniuk, Donoho, et al.
Kleinlet Compression

- This understanding of density can be applied to develop compression schemes.
- Earlier work, based on primary circle, called “Wedgelets”, done by Baraniuk, Donoho, et al.
- Extension to Klein bottle dictionary of patches natural.
Kleinlet Compression

A Picture is worth 1,000 words

The evidence for Kleinlets over Wedglets

Original

Coded by Kleinlet at .71bpp
PSNR= 29dB

Coded by Wedgelet at .8bpp
PSNR= 27.7dB

Kleinlet

Wedgelet

Kleinlet

Wedgelet
Kleinlet Compression

PSNR Comparisons

Kleinlets

Wedges

16x16 patches on a 512x512 image

PSNR=24.4

PSNR=22.9
Kleinlet Compression

Compression comparison between kleinlets and wedgelets

Cameraman

PSNR (dB) vs. Bits Per Pixel

Kleinlet: 29.2
Wedgelet: 28.8

1.4 dB
0.75 dB
Texture Recognition

- Texture patches can be sampled for high contrast patches
Texture Recognition

- Texture patches can be sampled for high contrast patches
- Yields distribution on Klein bottle
Texture Recognition

- Klein bottle has a natural geometry, and supports its own Fourier Analysis
Texture Recognition

- Klein bottle has a natural geometry, and supports its own Fourier Analysis
- Textures provide distributions on the Klein bottle
Texture Recognition

- Klein bottle has a natural geometry, and supports its own Fourier Analysis
- Textures provide distributions on the Klein bottle
- Pdf’s can be given Fourier expansions, gives coordinates for texture patches (Jose Perea)
Texture Recognition

- Klein bottle has a natural geometry, and supports its own Fourier Analysis
- Textures provide distributions on the Klein bottle
- Pdf’s can be given Fourier expansions, gives coordinates for texture patches (Jose Perea)
- Gives methods comparable to state of the art in performance, but in which effect of transformations such as rotation is predictable
Texture Recognition

Jose Perea - Duke University

Klein Bottle and Texture Discrimination
Summary

- Compression and texture recognition often obtained by using finite dictionaries
Compression and texture recognition often obtained by using finite dictionaries.

Geometry gives alternate notions of “finiteness”, i.e finite geometric descriptions of finite sets.
Compression and texture recognition often obtained by using finite dictionaries.

Geometry gives alternate notions of “finiteness”, i.e. finite geometric descriptions of finite sets.

Permits analysis using more mathematics, in particular coordinate changes.
Evolution

- Phylogenetics studies sets of sequences of various classes of organisms
Evolution

- Phylogenetics studies sets of sequences of various classes of organisms
- Uses Hamming or weighted versions of Hamming distances as organizing principle

Is this always justified?
Evolution

- Phylogenetics studies sets of sequences of various classes of organisms
- Uses Hamming or weighted versions of Hamming distances as organizing principle
- Often analyze by finding best approximation to space by trees
Evolution

- Phylogenetics studies sets of sequences of various classes of organisms
- Uses Hamming or weighted versions of Hamming distances as organizing principle
- Often analyze by finding best approximation to space by trees
- Is this always justified?
Theorem: Let $T$ be a tree, perhaps with lengths assigned to the edges. Then for any finite subspace of $T$, the persistent homology vanishes for every $i > 0$. This means there are no bars in higher degrees.
Evolution

Barcodes indicating the presence of “horizontal evolution”
Evolution

- Can study persistence barcodes of metric spaces of trees arising in evolution
Evolution

- Can study persistence barcodes of metric spaces of trees arising in evolution
- Presence of large loops can suggest that the standard model is incomplete

Evolution

- Can study persistence barcodes of metric spaces of trees arising in evolution
- Presence of large loops can suggest standard model is incomplete
- Signal of presence of alternate mechanisms, such as horizontal gene transfer

Evolution

- Can study persistence barcodes of metric spaces of trees arising in evolution
- Presence of large loops can suggest standard model is incomplete
- Signal of presence of alternate mechanisms, such as horizontal gene transfer
- Can also estimate various rates from the barcodes, by performing simulations

Evolution

- Can study persistence barcodes of metric spaces of trees arising in evolution
- Presence of large loops can suggest the standard model is incomplete
- Signal of presence of alternate mechanisms, such as horizontal gene transfer
- Can also estimate various rates from the barcodes by performing simulations
Other Applications of Persistence

- We have seen applications of persistent homology to individual data sets
Other Applications of Persistence

- We have seen applications of persistent homology to individual data sets
- Many times one has databases consisting of elements which themselves carry a metric space structure
Other Applications of Persistence

- We have seen applications of persistent homology to individual data sets
- Many times one has databases consisting of elements which themselves carry a metric space structure
- Molecules, images, ...
Other Applications of Persistence

- We have seen applications of persistent homology to individual data sets
- Many times one has databases consisting of elements which themselves carry a metric space structure
- Molecules, images, ...
- Can attach a barcode to each object
Other Applications of Persistence

- We have seen applications of persistent homology to individual data sets.
- Many times one has databases consisting of elements which themselves carry a metric space structure.
- Molecules, images, ...
- Can attach a barcode to each object.
- Gives a “non-linear indexing scheme” for such “unstructured” data.
Other Applications of Persistence

- We have seen applications of persistent homology to individual data sets.
- Many times one has databases consisting of elements which themselves carry a metric space structure.
- Molecules, images, ...
- Can attach a barcode to each object.
- Gives a “non-linear indexing scheme” for such “unstructured” data.
- Now one wants structures on space of barcodes for e.g. Machine Learning.
Other Applications of Persistence

- Barcodes form a metric space $\mathcal{B}$ under “bottleneck distance”
Barcodes form a metric space $\mathcal{B}$ under “bottleneck distance”

Each barcode $\beta_k(-)$ is Lipschitz with constant one from metric spaces with Gromov-Hausdorff metric to $\mathcal{B}$

Feature generation for this kind of data
Barcodes form a metric space $\mathcal{B}$ under “bottleneck distance”.
Each barcode $\beta_k(\cdot)$ is Lipschitz with constant one from metric spaces with Gromov-Hausdorff metric to $\mathcal{B}$.
$\mathcal{B}$ is also an infinite algebraic variety, suitably defined.
Other Applications of Persistence

- Barcodes form a metric space $\mathcal{B}$ under “bottleneck distance”
- Each barcode $\beta_k(-)$ is Lipschitz with constant one from metric spaces with Gromov-Hausdorff metric to $\mathcal{B}$
- $\mathcal{B}$ is also an infinite algebraic variety, suitably defined
- One obtains an infinite coordinatization of $\mathcal{B}$ using functions $\xi_{ij}$, $i > 0, j \geq 0$
Other Applications of Persistence

- Barcodes form a metric space $\mathcal{B}$ under “bottleneck distance”
- Each barcode $\beta_k(\cdot)$ is Lipschitz with constant one from metric spaces with Gromov-Hausdorff metric to $\mathcal{B}$
- $\mathcal{B}$ is also an infinite algebraic variety, suitably defined
- One obtains an infinite coordinatization of $\mathcal{B}$ using functions $\xi_{ij}, i > 0, j \geq 0$
- *Feature generation* for this kind of data
Thank you!