Topological Methods for Large and Complex Data Sets
IMA Workshop on Machine Learning, Minneapolis

Gunnar Carlsson, Stanford University

March 26, 2012
Shape of Data

- Data has shape
Shape of Data

- Data has shape
- The shape matters
Shape of Data

Clusters
Shape of Data

Predatory-Prey model
Shape of Data
Shape of Data

Flares
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

Log-log plot of a circle in the plane
Topology

Compressed Representations of Geometry
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 \( \mathbb{X} \), and a covering \( \mathcal{U} \).
Mapping

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

Build simplicial complex same way, but components replaced by clusters.
Mapping

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) $f : X \to Z$, 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 $Z$ - $\mathbb{R}$, $\mathbb{R}^2$, $S^1$.

The reference space typically has useful families of coverings attached to it.
Mapping
Mapping

Typical one dimensional filters:

- Density estimators
Mapping

Typical one dimensional filters:

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

Miller-Reaven Diabetes Dataset
Mapping

Cell Cycle Microarray Data

Joint with M. Nicolau, Nagarajan, G. Singh
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
Different platforms - importance of coordinate free approach
Mapping

Topological structure of leukemia

Data: Gene expression profiles of bone marrow of leukemia patients
Source: PMID 8573112
Columns: 1500 genes
Rows: 1905 patients
Mapping

Serendipity - copy number variation reveals parent child relations
Mapping

Netflix customer data
Mapping

Netflix customer data

“Bipolar” customers

“Negative” customers

One rating customers

“Positive” customers
Measuring Shape

$b_i$ is the “$i$-th Betti number”

$b_1=1$
$b_2=0$

$b_1=0$
$b_2=1$

$b_1=2$
$b_2=1$
Measuring Shape

Counts the number of “$i$-dimensional holes”

\begin{align*}
b_1 &= 1 \\
b_2 &= 0 \quad b_1 &= 0 \\
b_2 &= 1 \quad b_1 &= 2 \\
b_2 &= 1
\end{align*}
Measuring Shape

- 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 \( f: X \rightarrow Y \) induces linear transformation \( H_i(f): H_i(X) \rightarrow H_i(Y) \)
- 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 \textit{persistent homology}
- Developed by Edelsbrunner, Letscher, and Zomorodian and Zomorodian-Carlsson
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 suggest an approach
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 $\varepsilon$
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

$$V_0 \rightarrow V_1 \rightarrow V_2 \rightarrow V_3 \rightarrow \cdots$$
Measuring Shape of Data

- Provides an increasing sequence of simplicial complexes
- Apply $H_i$
- Gives a diagram of vector spaces
  \[ V_0 \rightarrow V_1 \rightarrow V_2 \rightarrow V_3 \rightarrow \cdots \]
- Call such algebraic structures *persistence vector spaces*
Measuring the Shape of Data

Can we classify persistence vector spaces, up to isomorphism?
Measuring the Shape of Data

- Transformation shows that classification of persistence vector spaces is equivalent to modules over $k[t]$
Measuring the Shape of Data

- Transformation shows that classification of persistence vector spaces is equivalent to modules over $k[t]$
- $k[t]$ is a PID
Measuring the Shape of Data

- Transformation shows that classification of persistence vector spaces is equivalent to modules over $k[t]$
- $k[t]$ is a PID
- So, module classification is as a direct sum of cyclic module
Let $[a, b]$ denote a persistence vector space with $V_i \cong k$ for $a \leq i \leq b$, with $V_i = \{0\}$ for $i \notin [a, b]$, and with all the maps between non-zero modules being identities. Not $b$ can be $+\infty$.

Every persistence vector space can be written uniquely as sums of modules of form $[a_i, b_i]$, so the isomorphism classes are parametrized by sets of intervals $[a_i, b_i]$. Call such a set a bar code. Long bars correspond to actual features, short ones to noise.
Measuring the Shape of Data

- Let $[a, b]$ denote a persistence vector space with $V_i \cong k$ for $a \leq i \leq b$, with $V_i = \{0\}$ for $i \notin [a, b]$, and with all the maps between non-zero modules being identities. Not $b$ can be $+\infty$.

- Every persistence vector space can be written uniquely as sums of modules of form $[a_i, b_i]$, so the isomorphism classes are parametrized by sets of intervals $[a_i, b_i]$. 
Measuring the Shape of Data

- Let \([a, b]\) denote a persistence vector space with \(V_i \cong k\) for \(a \leq i \leq b\), with \(V_i = \{0\}\) for \(i \notin [a, b]\), and with all the maps between non-zero modules being identities. Not \(b\) can be \(+\infty\).

- Every persistence vector space can be written uniquely as sums of modules of form \([a_i, b_i]\), so the isomorphism classes are parametrized by sets of intervals \([a_i, b_i]\).

- Call such a set a *bar code*. 


Measuring the Shape of Data

- Let \([a, b]\) denote a persistence vector space with \(V_i \cong k\) for \(a \leq i \leq b\), with \(V_i = \{0\}\) for \(i \not\in [a, b]\), and with all the maps between non-zero modules being identities. Not \(b\) can be \(+\infty\).

- Every persistence vector space can be written uniquely as sums of modules of form \([a_i, b_i]\), so the isomorphism classes are parametrized by sets of intervals \([a_i, b_i]\).

- Call such a set a bar code.

- Long bars correspond to actual features, short ones to noise.
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.
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 *pixel space*, $\mathcal{P}$. 
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?
Natural Images

1. $\mathcal{I}$ is very high dimensional, because images are so expressive
1. $\mathcal{I}$ is very high dimensional, because images are so expressive

2. $\mathcal{I}$ is very sparse (of high codimension) in $\mathcal{P}$, because a random choice of values for each pixel will not give anything close to an image
Natural Images

1. $\mathcal{I}$ is very high dimensional, because images are so expressive.

2. $\mathcal{I}$ is very sparse (of high codimension) in $\mathcal{P}$, because a random choice of values for each pixel will not give anything close to an image.

3. The direct study of $\mathcal{I}$ will be very difficult, unless one severely restricts the range of subjects for the images.
Natural Images

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

Observations:
Patches

Observations:

1. Each patch gives a vector in $\mathbb{R}^9$
Patches

Observations:

1. Each patch gives a vector in $\mathbb{R}^9$
2. Most patches will be nearly constant, or low contrast, because of the presence of regions of solid shading in most images
Observations:

1. Each patch gives a vector in \( \mathbb{R}^9 \)

2. Most patches will be nearly constant, or \textit{low contrast}, because of the presence of regions of solid shading in most images

3. Low contrast will dominate statistics, not interesting
Lee-Mumford-Pedersen [LMP] study only high contrast patches
Patches

- Lee-Mumford-Pedersen [LMP] study only high contrast patches

- Contrast measured by quadratic form on space of patches, so-called *D-norm*
Patches

- Lee-Mumford-Pedersen [LMP] study only high contrast patches

- Contrast measured by quadratic form on space of patches, so-called $D$-norm

- Restricts study to parts of the images where transitions in intensity occur, i.e. edges, lines, etc.
Patches

- Lee-Mumford-Pedersen [LMP] study only high contrast patches

- Contrast measured by quadratic form on space of patches, so-called $D$-norm

- Restricts study to parts of the images where transitions in intensity occur, i.e. edges, lines, etc.

- How do they study the high contrast patches?
Patches

- Collect c:a $4.5 \times 10^6$ high contrast patches from a collection of images obtained by van Hateren and van der Schaaf
Patches

- Collect ca $4.5 \times 10^6$ high contrast patches from a collection of images obtained by van Hateren and van der Schaar.

- Normalize mean intensity by subtracting mean from each pixel value to obtain patches with mean intensity $= 0$. 

[Image]
Patches

- Collect c:a $4.5 \times 10^6$ high contrast patches from a collection of images obtained by van Hateren and van der Schaaf.

- Normalize mean intensity by subtracting mean from each pixel value to obtain patches with mean intensity $= 0$.

- Puts data on an 8-dimensional hyperplane, $\cong \mathbb{R}^8$. 
Patches

- Collect c:a $4.5 \times 10^6$ high contrast patches from a collection of images obtained by van Hateren and van der Schaaf.

- Normalize mean intensity by subtracting mean from each pixel value to obtain patches with mean intensity $= 0$.

- Puts data on an 8-dimensional hyperplane, $\cong \mathbb{R}^8$.

- Means that we will consider as equivalent patches which can be obtained from each other by turning the intensity knob.
Patches

- Normalize contrast by dividing by the $D$-norm, so obtain patches with $D$-norm $= 1$
Patches

- Normalize contrast by dividing by the $D$-norm, so obtain patches with $D$-norm = 1

- Means that data now lies on a 7-D ellipsoid, $\cong S^7$
Patches

- Normalize contrast by dividing by the $D$-norm, so obtain patches with $D$-norm = 1

- Means that data now lies on a 7-D ellipsoid, $\cong S^7$

- Normalization means that we will consider patches which can be obtained from each other by turning contrast knob to be the same
Patches
Result: Point cloud data $\mathcal{M}$ lying on a sphere in $\mathbb{R}^8$
Patches

**Result:** Point cloud data $\mathcal{M}$ lying on a sphere in $\mathbb{R}^8$

We wish to analyze it with persistent homology to understand it *qualitatively*.
Analysis

**First Observation:** The points fill out $S^7$ in the sense that every point in $S^7$ is “close” to a point in $M$. 
First Observation: The points fill out $S^7$ in the sense that every point in $S^7$ is “close” to a point in $\mathcal{M}$

Initially disappointing, since it means that nothing special can be said about the actual patches different from patches chosen at random
Analysis

**First Observation:** The points fill out $S^7$ in the sense that every point in $S^7$ is “close” to a point in $M$

Initially disappointing, since it means that nothing special can be said about the actual patches different from patches chosen at random

However, density of points varies a great deal from region to region
First Observation: The points fill out $S^7$ in the sense that every point in $S^7$ is “close” to a point in $M$.

Initially disappointing, since it means that nothing special can be said about the actual patches different from patches chosen at random.

However, density of points varies a great deal from region to region.

Study the subsets of high density, as measured by different density estimators.
Primary Circle

$5 \times 10^4$ points, $k = 300$, $T = 25$

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

$5 \times 10^4$ 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
Three Circle Model

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

\[ \beta_1 = 5 \]
Three Circle Model

Does the data fit with this model?
Three Circle Model

SECONDARY CIRCLE
Three Circle Model
Database

T = 5%

T = 25%
Three Circle Model

IS THERE A TWO DIMENSIONAL SURFACE IN WHICH THIS PICTURE FITS?
Klein Bottle

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

<table>
<thead>
<tr>
<th>Betti 0</th>
<th>Betti 1</th>
<th>Betti 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

Betti 0 = 1
Betti 1 = 2
Betti 2 = 1
Klein Bottle

$\mathcal{K} - KLEIN\ BOTTLE$
Klein Bottle

\begin{align*}
\begin{array}{c|ccc}
  i & 0 & 1 & 2 \\
\hline
\beta_i(K) & 1 & 2 & 1 \\
\end{array}
\end{align*}

Agrees with the Betti numbers we found from data.
Klein Bottle

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

![Klein Bottle Diagram](image-url)

**Primary Circle**

- Points: P, Q

Diagram shows a square with arrows indicating the path of the primary circle, starting at P, going around the square, and ending at Q.
Klein Bottle
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. \( \int_D f^2 = 1 \)
Applications of Persistence

- By using persistence on other quantities (density, centrality, ...) can get useful shape invariants

Can one do machine learning on barcode space?
Applications of Persistence

- By using persistence on other quantities (density, centrality, ...) can get useful shape invariants
- Persistence barcodes lie in barcode space, has a metric
- Can one do machine learning on barcode space?
Applications of Persistence

- By using persistence on other quantities (density, centrality, ...) can get useful shape invariants
- Persistence barcodes lie in *barcode space*, has a metric
- Persistence gives a map $\Psi$ from $\mathcal{M}$ (space of metric spaces with Gromov-Hausdorff metric) to $\mathcal{B}$ (barcode space with bottleneck distance)
Applications of Persistence

- By using persistence on other quantities (density, centrality, ...) can get useful shape invariants
- Persistence barcodes lie in \textit{barcode space}, has a metric
- Persistence gives a map $\mathcal{P}$ from $\mathcal{M}$ (space of metric spaces with Gromov-Hausdorff metric) to $\mathcal{B}$ (barcode space with bottleneck distance)
- $\mathcal{P}$ is distance non-increasing (Chazal, Mémoli, Guibas, Oudot)

Can one do machine learning on barcode space?
Applications of Persistence

- By using persistence on other quantities (density, centrality, ...) can get useful shape invariants
- Persistence barcodes lie in barcode space, has a metric
- Persistence gives a map $\mathcal{P}$ from $M$ (space of metric spaces with Gromov-Hausdorff metric) to $\mathcal{B}$ (barcode space with bottleneck distance)
- $\mathcal{P}$ is distance non-increasing (Chazal, Mémoli, Guibas, Oudot)
- Can be used to get useful invariants of shapes - X-ray images, for example
Applications of Persistence

- By using persistence on other quantities (density, centrality, ...) can get useful shape invariants
- Persistence barcodes lie in *barcode space*, has a metric
- Persistence gives a map $\mathcal{P}$ from $\mathcal{M}$ (space of metric spaces with Gromov-Hausdorff metric) to $\mathcal{B}$ (barcode space with bottleneck distance)
- $\mathcal{P}$ is distance non-increasing (Chazal, Mémoli, Guibas, Oudot)
- Can be used to get useful invariants of shapes - X-ray images, for example
- Can one do machine learning on barcode space?
Applications of Persistence

- Space of barcodes can be thought of as an “infinite algebraic variety”
Applications of Persistence

- Space of barcodes can be thought of as an “infinite algebraic variety”
- Get a ring of algebraic functions which detect barcodes
Applications of Persistence

- Space of barcodes can be thought of as an “infinite algebraic variety”
- Get a ring of algebraic functions which detect barcodes
- Ring analyzed by Adcock, E. Carlsson, G.C.