The Shape of Biomedical Data

ACM-BCB 2016

Gunnar Carlsson

Stanford University and Ayasdi Inc.

October 2, 2016

Big Data

Its 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

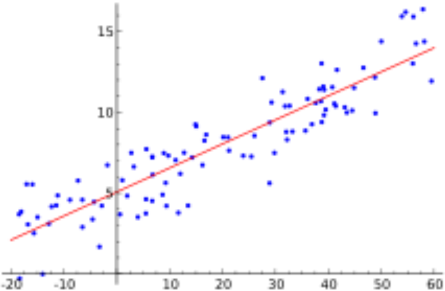
Shape of Data

- Data has shape

Shape of Data

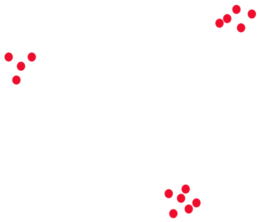
- Data has shape
- The shape matters

Shape of Data



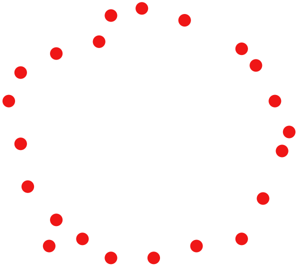
Linear Regression

Shape of Data



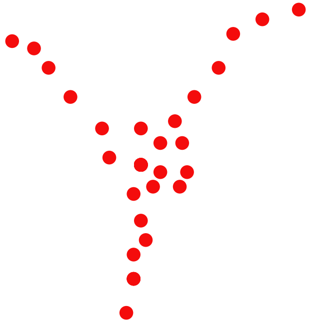
Clusters

Shape of Data



Loop

Shape of Data



“Y-junction”

Shape of Data

- How to model data?

Shape of 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

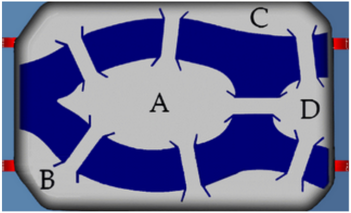
Topology

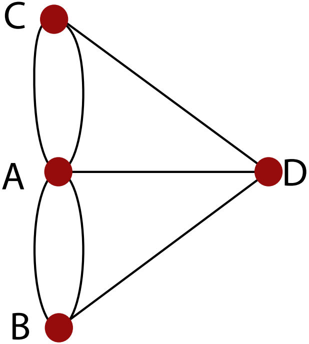
- Formalism for measuring and representing shape
- Pure mathematics since 1700's

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:

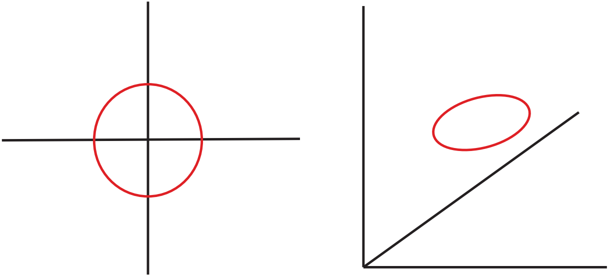
- Coordinate freeness
- Invariance under deformation

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

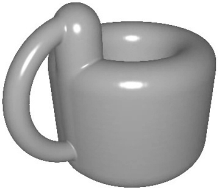
Topology

Coffee cup is the “same” as a doughnut

Topology

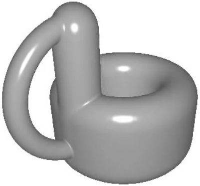
Coffee cup is the “same” as a doughnut

Topology



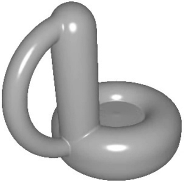
Coffee cup is the “same” as a doughnut

Topology



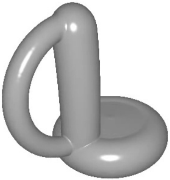
Coffee cup is the "same" as a doughnut

Topology



Coffee cup is the “same” as a doughnut

Topology

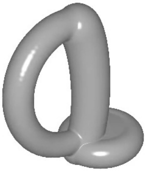


Coffee cup is the "same" as a doughnut

Topology

Coffee cup is the "same" as a doughnut

Topology



Coffee cup is the “same” as a doughnut

Topology

Coffee cup is the "same" as a doughnut

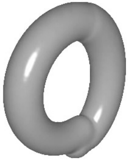
Topology

Coffee cup is the "same" as a doughnut

Topology

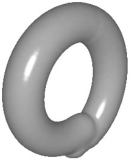
Coffee cup is the "same" as a doughnut

Topology



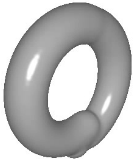
Coffee cup is the “same” as a doughnut

Topology



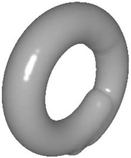
Coffee cup is the "same" as a doughnut

Topology



Coffee cup is the “same” as a doughnut

Topology

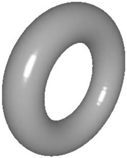


Coffee cup is the "same" as a doughnut

Topology

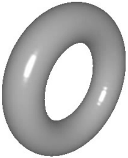
Coffee cup is the “same” as a doughnut

Topology



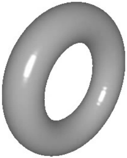
Coffee cup is the "same" as a doughnut

Topology



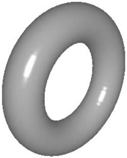
Coffee cup is the "same" as a doughnut

Topology



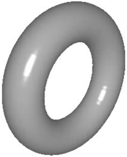
Coffee cup is the "same" as a doughnut

Topology



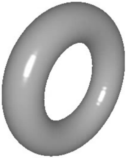
Coffee cup is the “same” as a doughnut

Topology



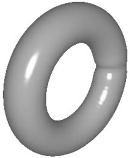
Coffee cup is the "same" as a doughnut

Topology



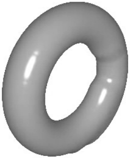
Coffee cup is the "same" as a doughnut

Topology



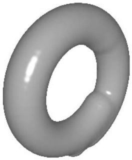
Coffee cup is the "same" as a doughnut

Topology



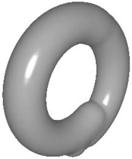
Coffee cup is the "same" as a doughnut

Topology



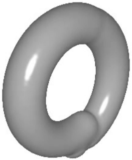
Coffee cup is the “same” as a doughnut

Topology



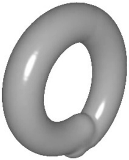
Coffee cup is the "same" as a doughnut

Topology



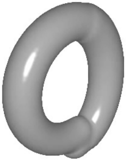
Coffee cup is the “same” as a doughnut

Topology



Coffee cup is the “same” as a doughnut

Topology

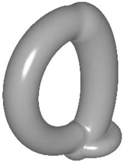


Coffee cup is the "same" as a doughnut

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

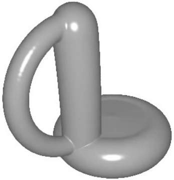
Topology

Coffee cup is the “same” as a doughnut

Topology

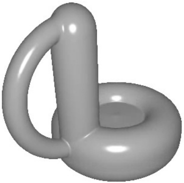
Coffee cup is the "same" as a doughnut

Topology



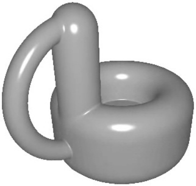
Coffee cup is the “same” as a doughnut

Topology



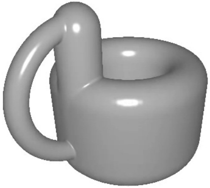
Coffee cup is the "same" as a doughnut

Topology



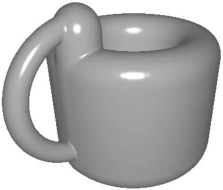
Coffee cup is the “same” as a doughnut

Topology



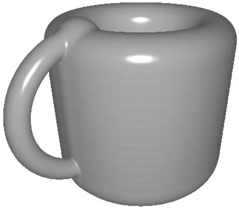
Coffee cup is the “same” as a doughnut

Topology



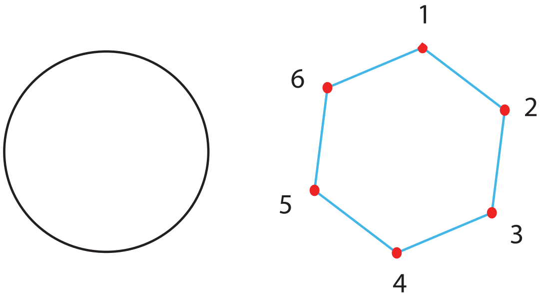
Coffee cup is the "same" as a doughnut

Topology



Coffee cup is the "same" as a doughnut

Topology



Compressed Representations of Geometry

Topology

Two tasks:

Topology

Two tasks:

- Represent shape

Topology

Two tasks:

- Represent shape
- Measure shape

Representing Shape

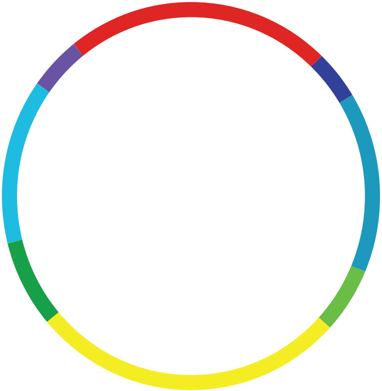
Can one extend topological mapping methods (compressed representations) from idealized shapes to data?

Representing Shape

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

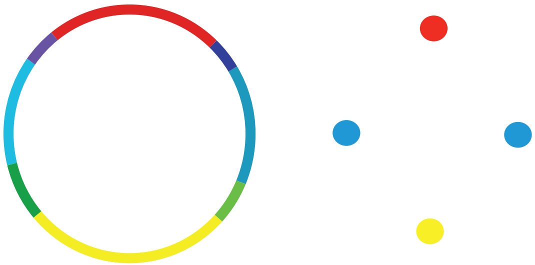
Yes (Singh, Memoli, G. C.)

Topological Mapping



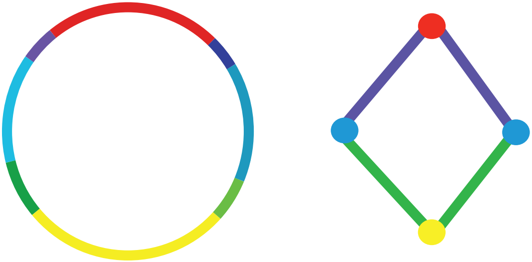
Covering of Circle

Topological Mapping



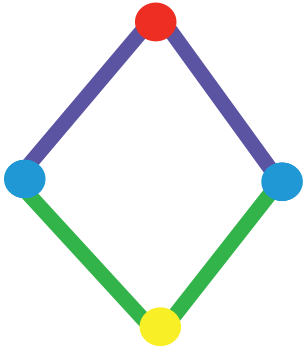
Create nodes

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?

Mapping

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.

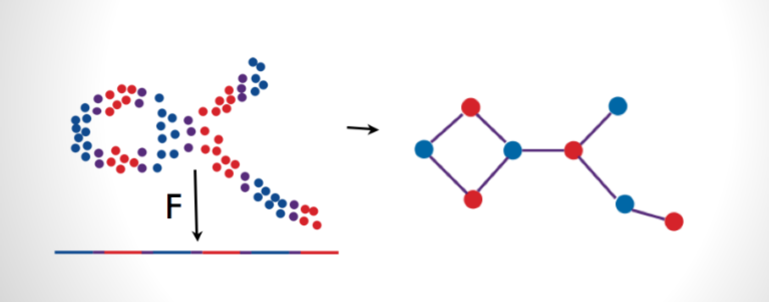
Mapping

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



Mapping

Typical one dimensional filters:

- Density estimators

Mapping

Typical one dimensional filters:

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

Mapping

Typical one dimensional filters:

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

Mapping

Typical one dimensional filters:

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

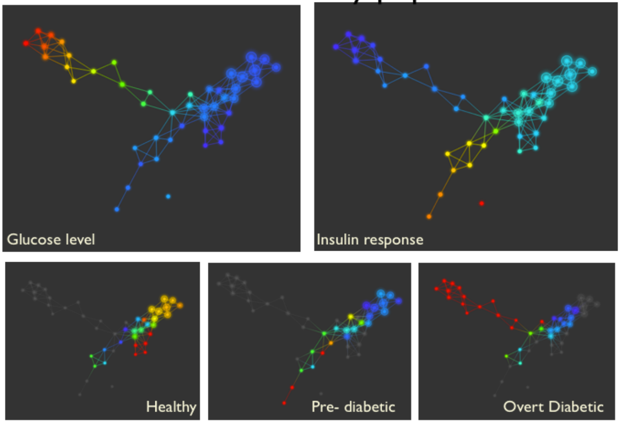
Mapping

Typical one dimensional filters:

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

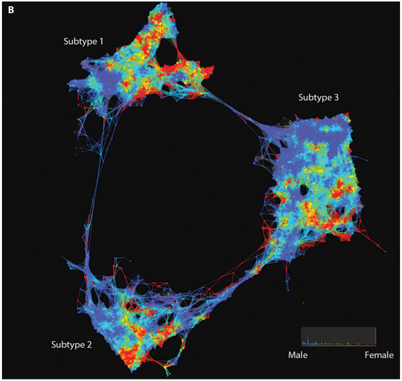
Mapping

Relationships between diabetic, pre-diabetic and healthy populations



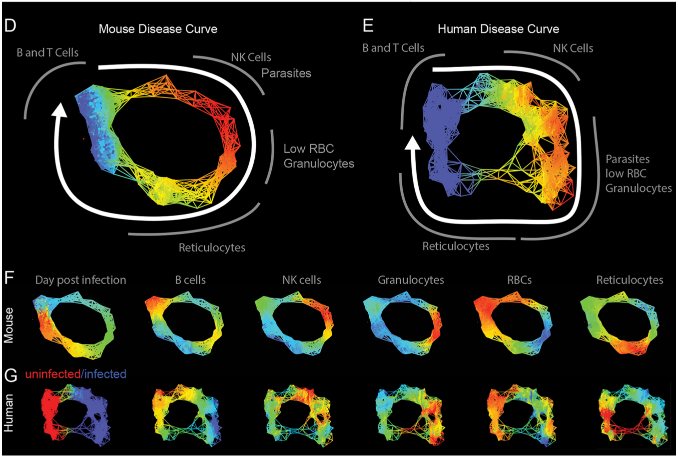
Miller-Reaven Diabetes Dataset

Mapping



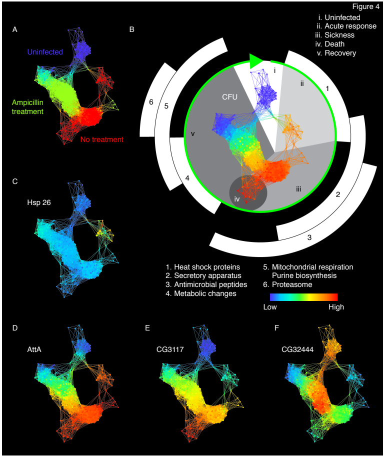
Li et al, Science Translational Medicine, 2015

Mapping



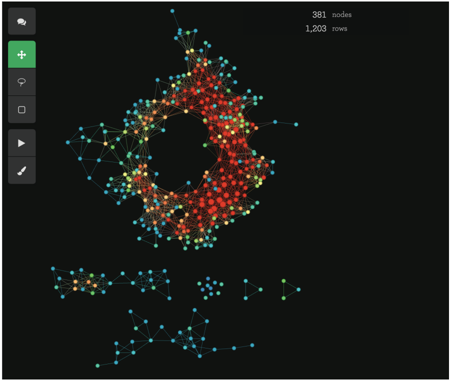
Torres et al, PLOS Biology, 2016

Mapping



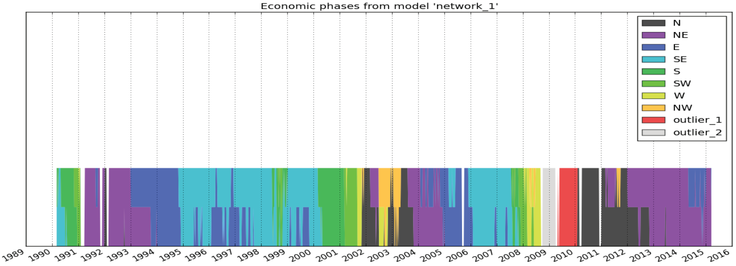
Louie et al, PLOS Biology, 2016

Mapping



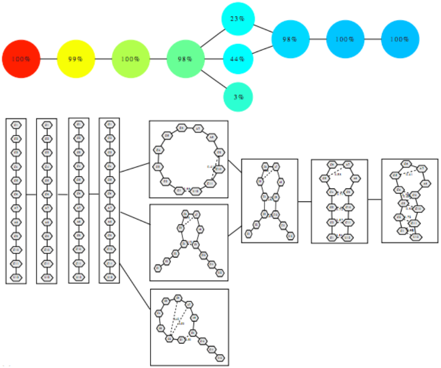
Economic Regime Analysis

Mapping



Economic Regime Analysis

Mapping



RNA hairpin folding data
Joint with G. Bowman, X. Huang, Y. Yao, J. Sun, L. Guibas, V. Pande, J. Chem. Physics, 2009

Mapping

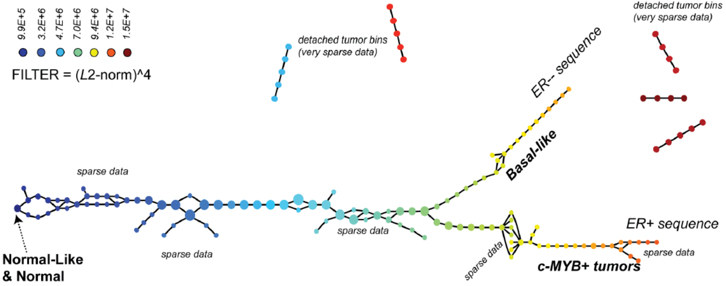
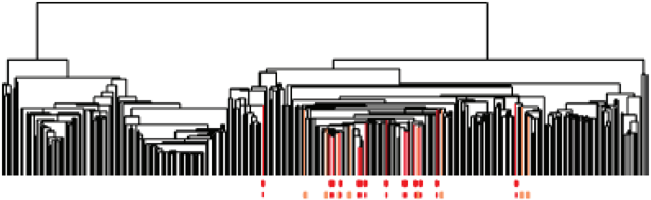
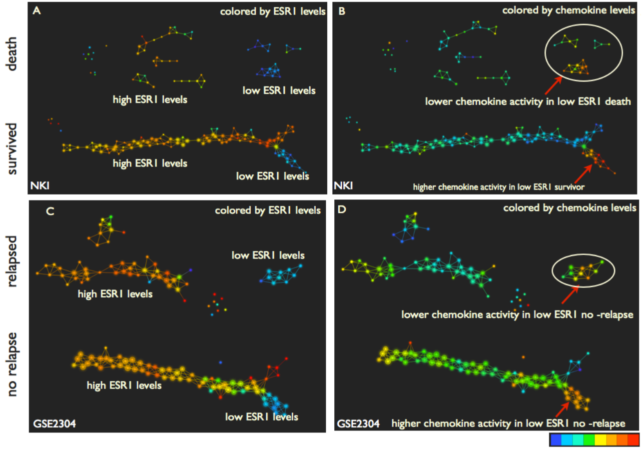


Diagram of gene expression profiles for breast cancer
M. Nicolau, A. Levine, and G. Carlsson, PNAS 2011

Mapping

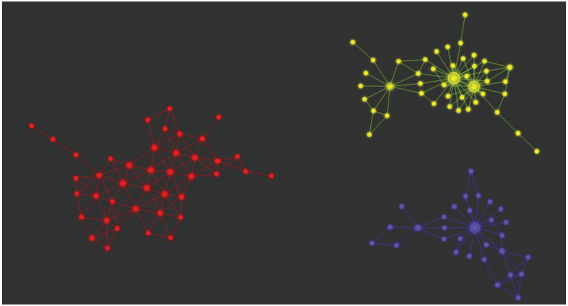


Comparison with hierarchical clustering



Different platforms - importance of coordinate free approach

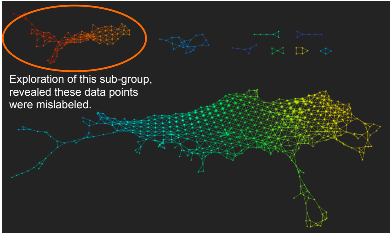
Mapping



Serendipity - copy number variation reveals parent child relations

Mapping

Example: Quality Control



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

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.

AYASDI
Discover what you don't know.

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

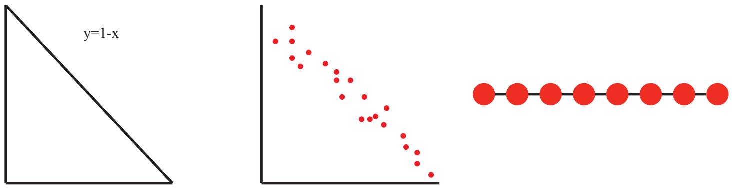
Topological Modeling

- 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

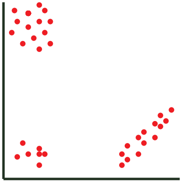
- Suggests a new kind of modeling
- Output is no longer a set of algebraic formulae, but a network
- Input is a finite set equipped with a distance function
- Distance function encodes similarity

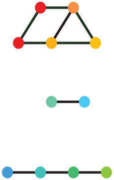
Topological Modeling



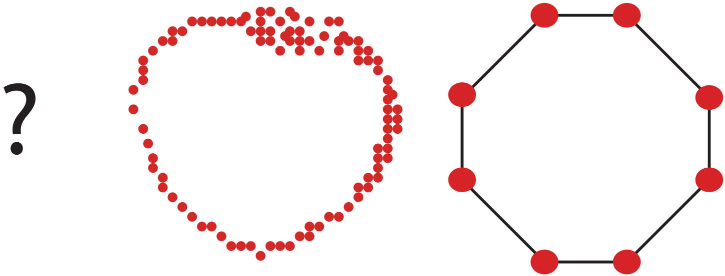
Topological Modeling

A	B	C	D	E
Group 1	100	5.123	17.89	0.05589715
	101	199.75	36.52	0.02738226
	102	201.75	73.78	0.01355381
	103	203.75	148.3	0.00674309
	104	205.75	297.34	0.00336315
	105	207.75	595.42	0.00167949
	106	209.75	1191.58	0.00083922
	107	211.75	2383.9	0.00041948
	108	213.75	4768.54	0.00020971
	109	215.75	9537.82	0.00010485
Group 2	205	217.75	19076.38	0.00457143
	208	409.75	38153.5	0.00243457
	211	415.75	76307.74	0.00239952
	214	421.75	152616.22	0.00236546
	217	427.75	305233.18	0.00233236
	220	433.75	610467.1	0.00230017
	223	439.75	1220934.94	0.00226886
	226	445.75	2441870.62	0.00223839
	229	451.75	4883741.98	0.00220872
Group 3	267	457.75	9767484.7	0.00217984
	269	533.75	19534970.1	0.00187003
	271	537.75	39069941	0.00185615
	273	541.75	78139882.8	0.00184247
	275	545.75	156279766	0.00182899
	277	549.75	312559533	0.00181571
	279	553.75	625119067	0.00180261
	281	557.75	1250238136	0.00178971
	283	561.75	2500476272	0.00177699
	285	565.75	5000952545	0.00176445
	287	569.75	1.0002E+10	0.00175208

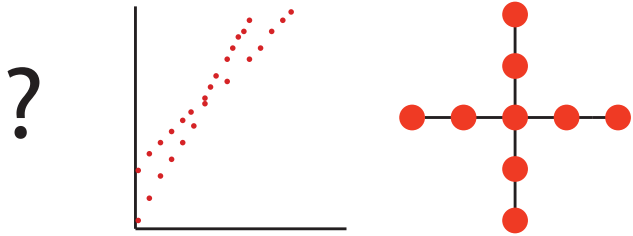




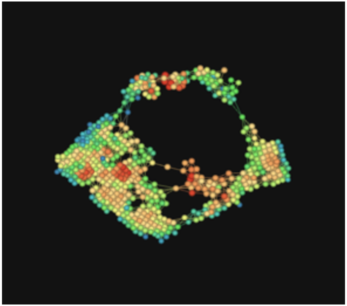
Topological Modeling



Topological Modeling

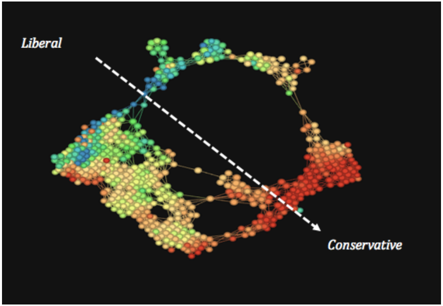


Topological Modeling - Coloring by Function Values



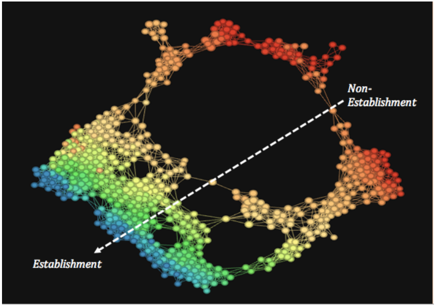
World Values Survey - 2000 U.S. Respondents
11 Questions on Trust in Institutions

Topological Modeling - Coloring by Function Values



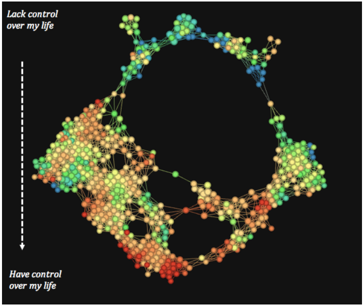
Color by response to left/right preference

Topological Modeling - Coloring by Function Values



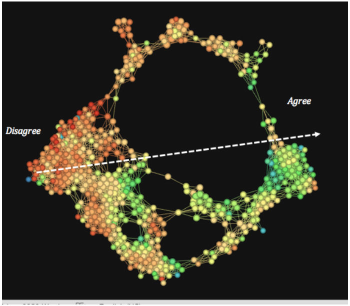
Coloring by sum of trust in all 11 institutions

Topological Modeling - Coloring by Function Values



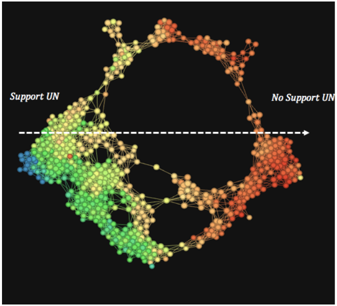
Color by response to "Do you feel you have control over your life?"

Topological Modeling - Coloring by Function Values



Response to “Should employers favor native born employees in difficult economic times?”

Topological Modeling - Coloring by Function Values



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

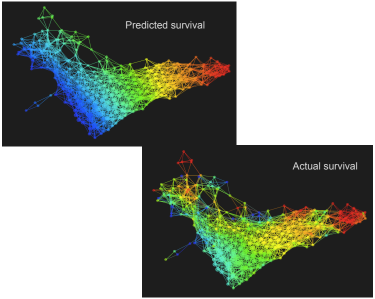
- Suppose we are given outcome of interest, such as "survival", "revenue", "fraud", "Democrat/Republican", etc.
- Coloring by average value of outcome on data points in node is useful
- Frequently discover "hot spots" of concentration of high values of the outcome

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
- Extremely useful information

Topological Modeling - Hot Spot Analysis

Example: Model Verification



About the Data

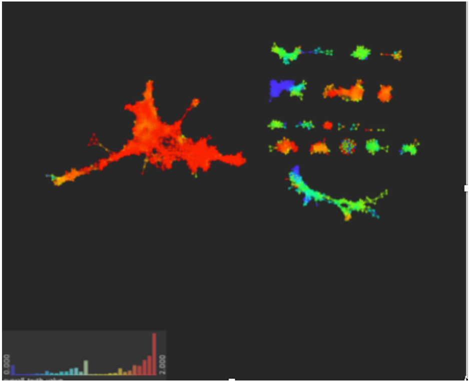
When patients come to an emergent care facility, doctors need to assess priority and predict probability of survival with medical intervention.

Patient is quickly assessed for information about their condition: temperature, blood pressure, yes/no questions.

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.

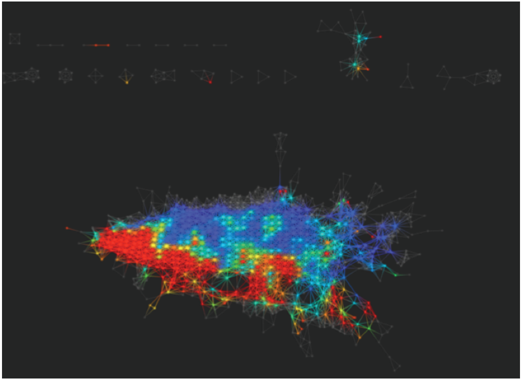
AYASDI
Discover what you don't know.

Topological Modeling - Hot Spot Analysis



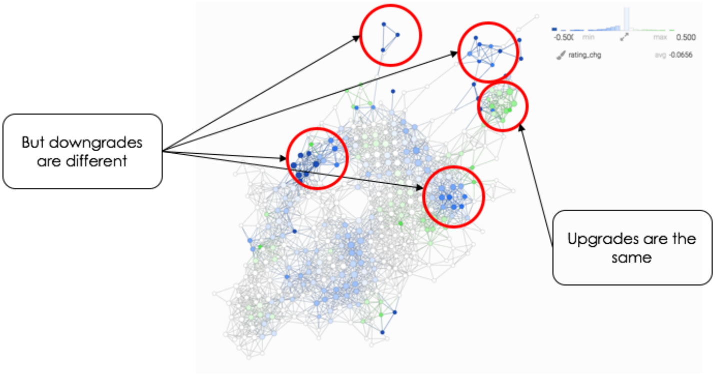
Program Downgrades

Topological Modeling - Hot Spot Analysis



Program Downgrades

Topological Modeling - Hot Spot Analysis



Credit Risk Analysis

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

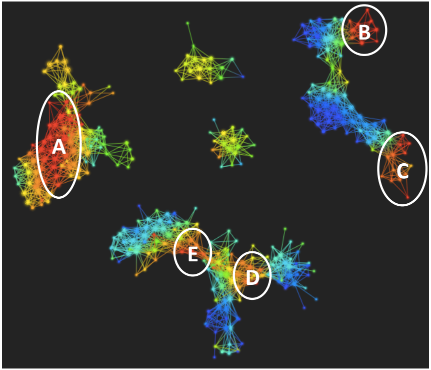
Topological Modeling - Feature Selection

- 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

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
- Hot spot analysis in columns is also useful

Topological Modleing - 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

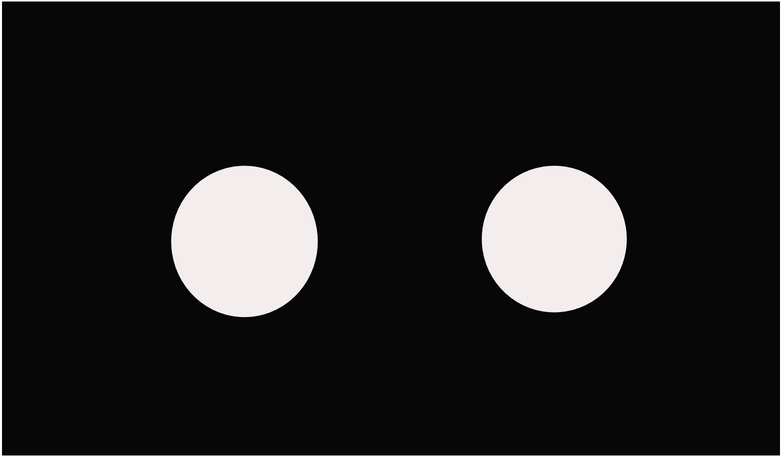
Measuring Shape

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

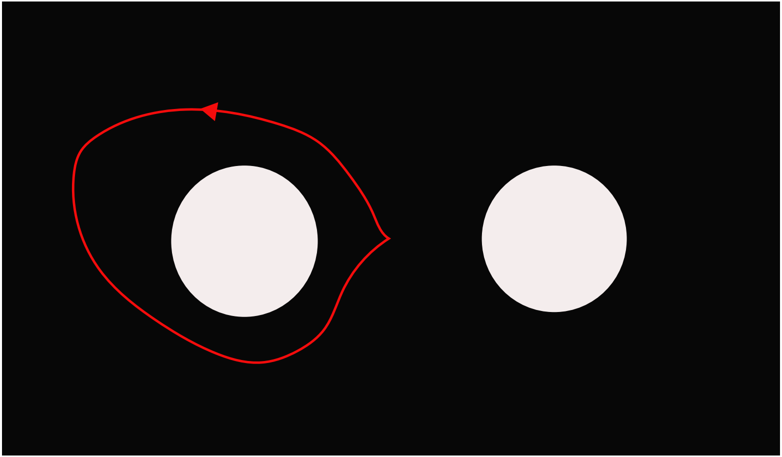
Measuring Shape

- 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 am appropriate sense

Measuring Shape

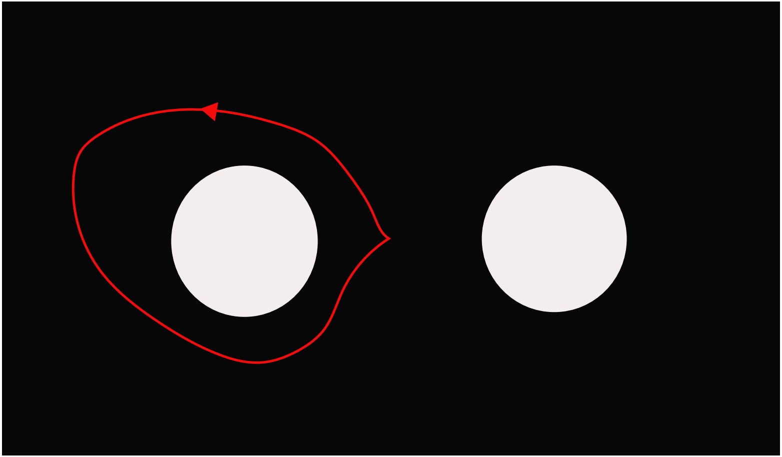


Measuring Shape



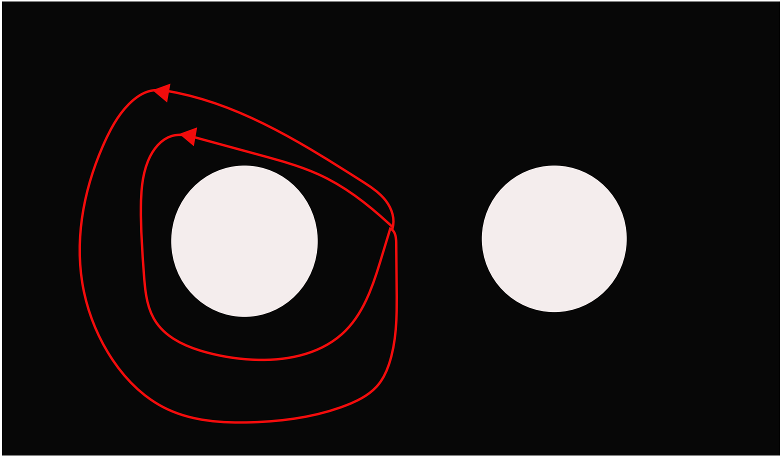
Capturing obstacle by "lassoing"

Measuring Shape



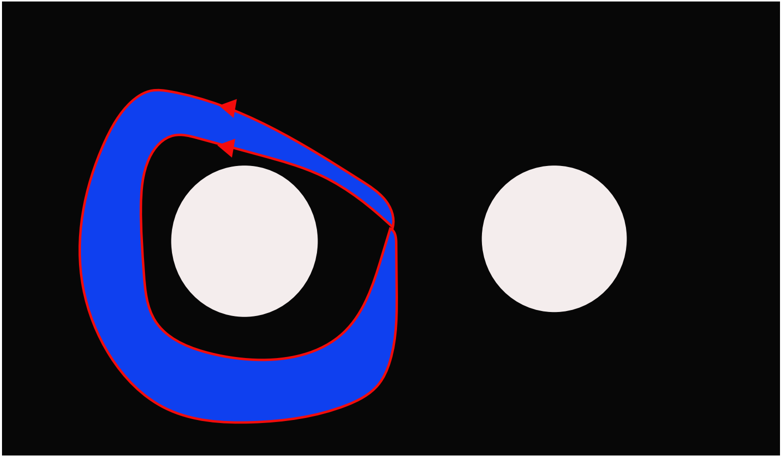
Capturing obstacle by "lassoing"

Measuring Shape



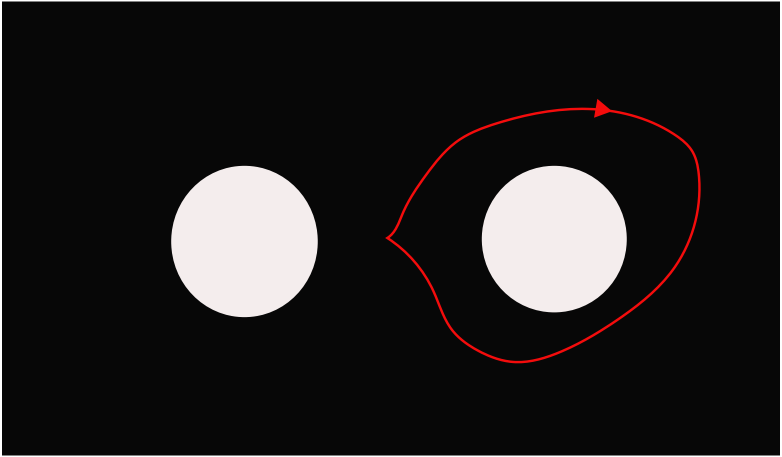
Two different lassos capture same obstacle

Measuring Shape



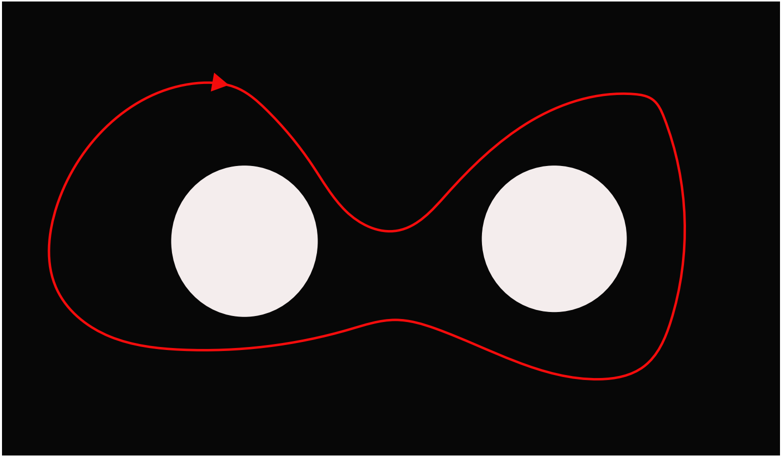
Solve by introducing homotopy relation

Measuring Shape



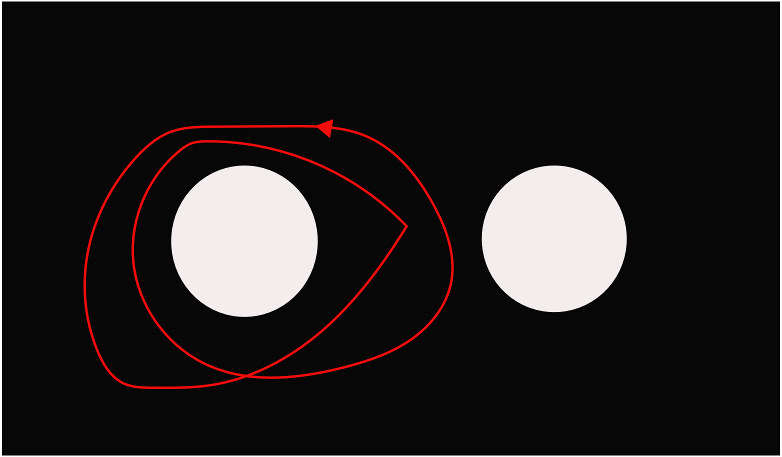
Second different lasso

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*

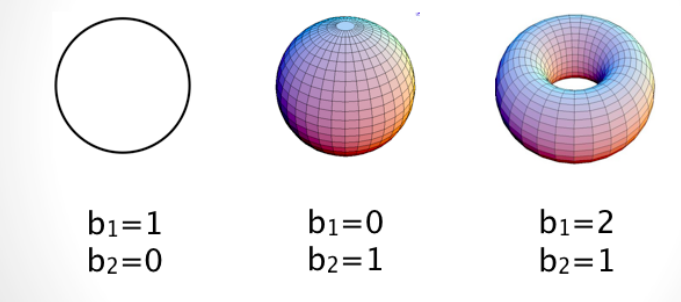
Measuring Shape

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

Measuring Shape

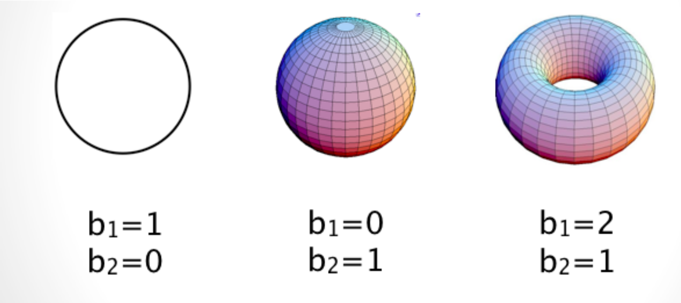
- Algebraic topology performs counts of occurrences of *equivalence classes of geometric patterns*
- Naive counting typically give infinite answers
- Counting is done by computing dimensions of algebraic objects

Measuring Shape



b_i is the "i-th Betti number"

Measuring Shape



Counts the number of "*i*-dimensional holes"

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)
- $b_i(X) = dimH_i(X)$
- $H_i(X)$ 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?

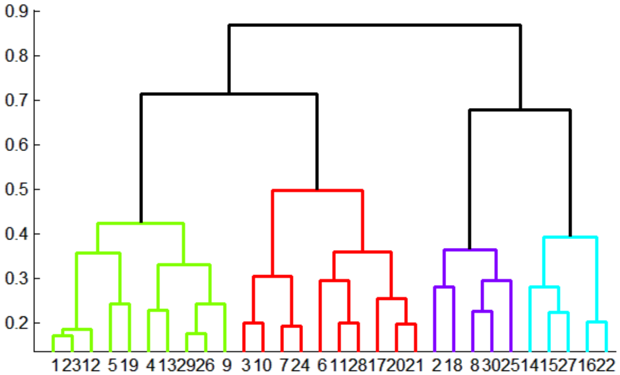
Measuring Shape of Data

- 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 threshhold ϵ

Measuring Shape of Data

- Points are connected when they are within a threshhold ϵ
- Dendrogram gives a profile of the clustering at all ϵ's simultaneously

Measuring Shape of Data

- Points are connected when they are within a threshhold ϵ
- Dendrogram gives a profile of the clustering at all ϵ's simultaneously
- Doesn't require choosing a threshhold

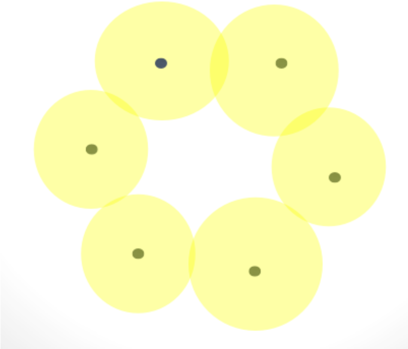
Measuring Shape of Data

- How to build spaces from finite metric spaces

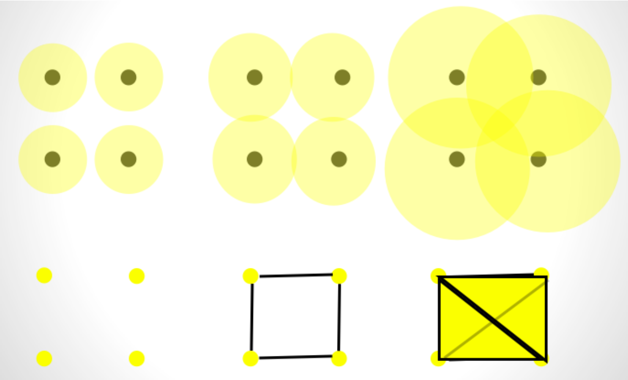
Measuring Shape of Data

- How to build spaces from finite metric spaces
- Use the nerve of the covering by balls of a given radius ϵ

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)

$$V_0 \to V_1 \to V_2 \to V_3 \to \cdots$$

- 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

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

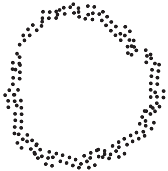
Measuring the Shape of Data

- 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

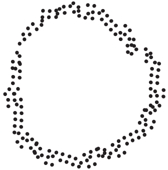
Measuring the Shape of Data

- 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

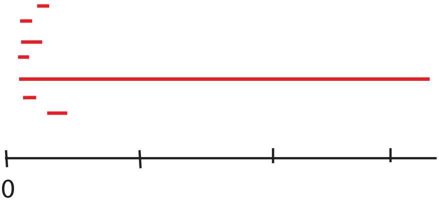
Measuring the Shape of Data - Barcodes



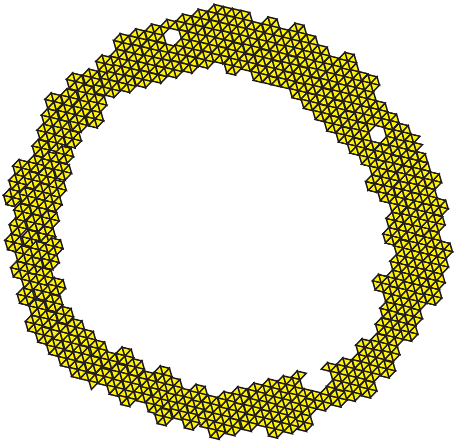
Measuring the Shape of Data - Barcodes



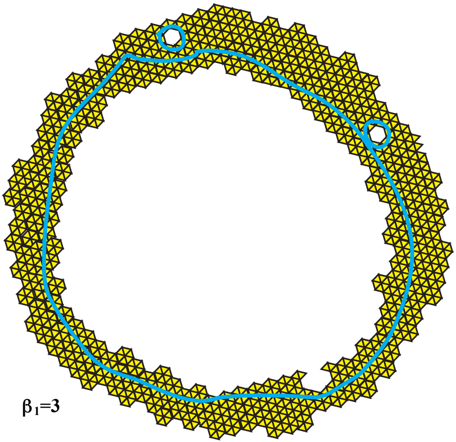
One dimensional barcode:



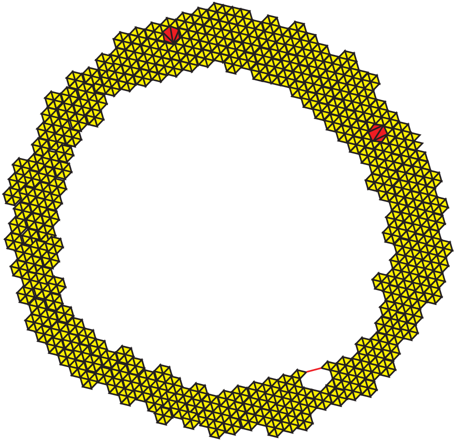
Measuring the Shape of Data - Barcodes



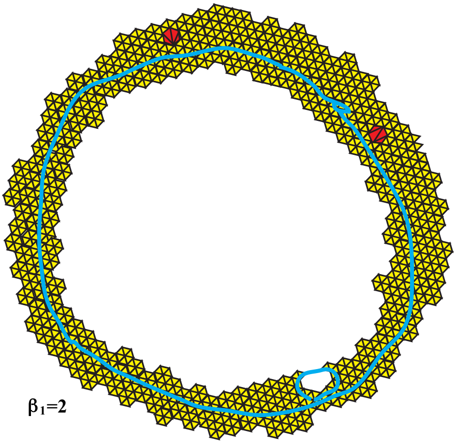
Measuring the Shape of Data - Barcodes



Measuring the Shape of Data - Barcodes



Measuring the Shape of Data - Barcodes



Application to Natural Image Statistics

With V. de Silva, T. Ishkanov, A. Zomorodian

Natural Images

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 *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

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 3×3 patches in the image.

Natural Images

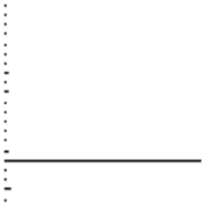
Solution (Lee, Mumford, Pedersen): Study *local* structure of images statistically, where there is less variation

Specifically, study 3×3 patches in the image.

Study high *density* high *contrast* patches

Primary Circle

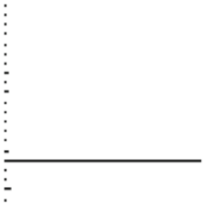
5×10^4 points, $k = 300$, $T = 25$



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

Primary Circle

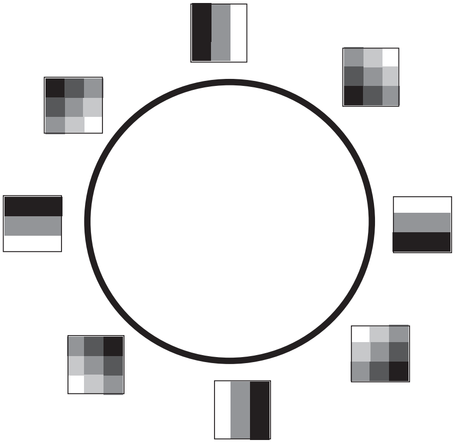
$$5 \times 10^4 \text{ points}, \; k = 300, \; T = 25$$



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

Is the set clustered around a circle?

Primary Circle



PRIMARY CIRCLE

Three Circle Model

5×10^4 points, $k = 15$, $T = 25$

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

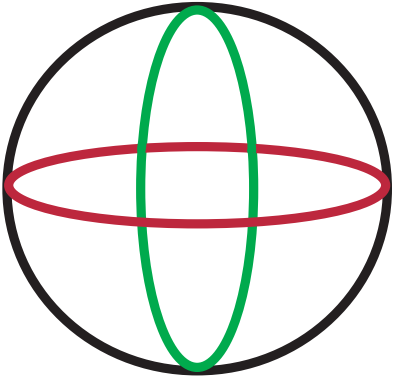
Three Circle Model

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

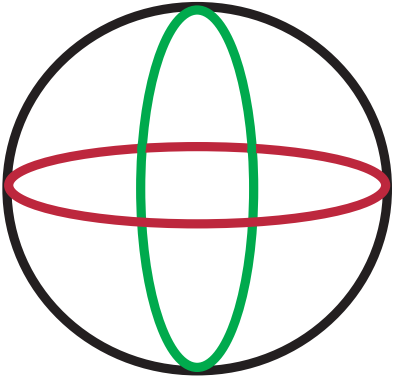
One-dimensional barcode, suggests $\beta_1 = 5$

What's the explanation for this?

Three Circle Model

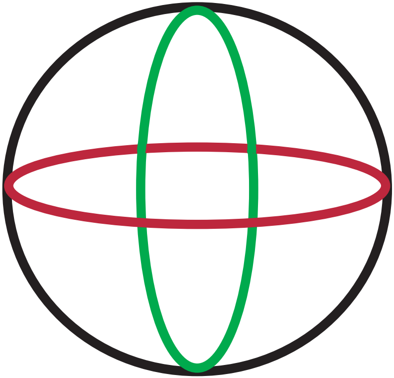


Three Circle Model



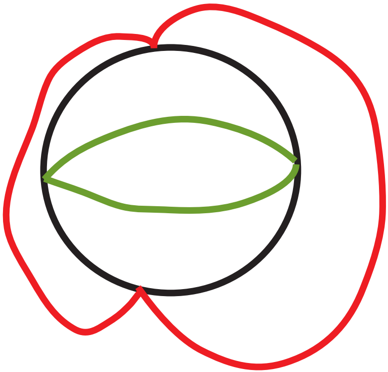
THREE CIRCLE MODEL

Three Circle Model

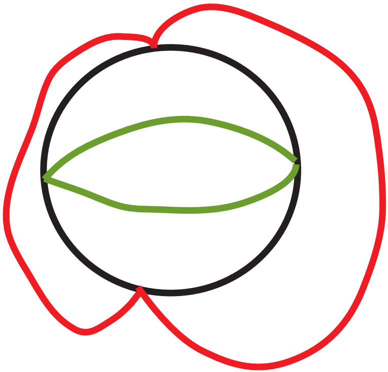


Red and green circles do not touch, each touches black circle

Three Circle Model

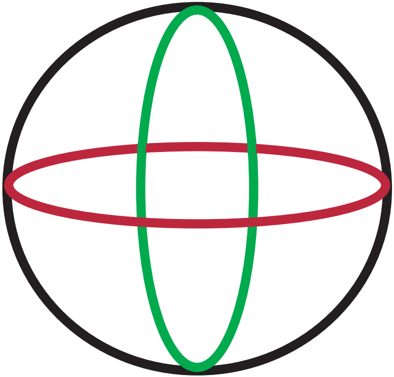


Three Circle Model



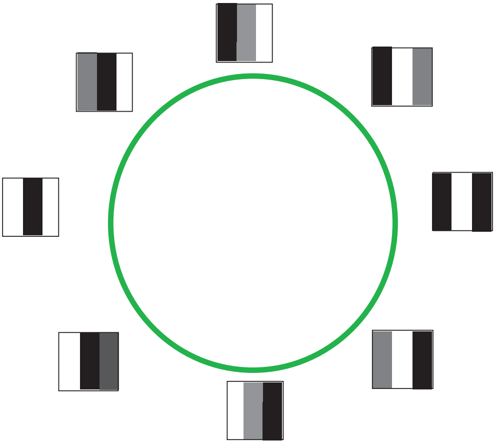
$\beta_1 = 5$

Three Circle Model



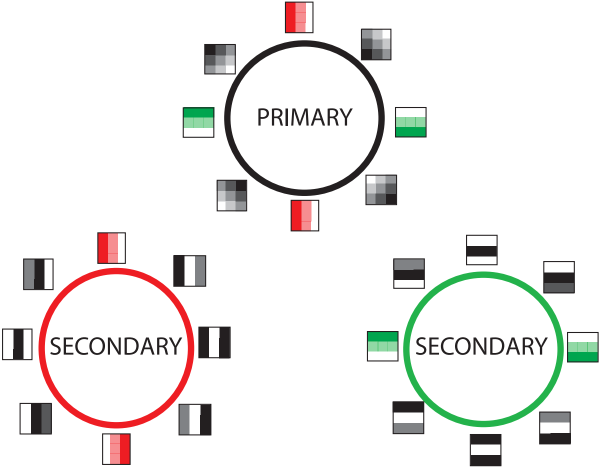
Does the data fit with this model?

Three Circle Model

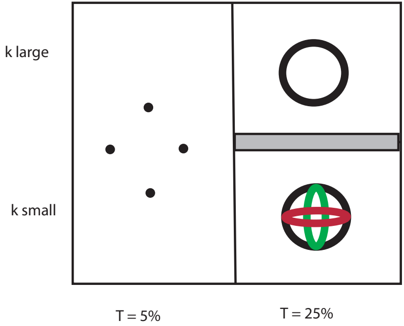


SECONDARY CIRCLE

Three Circle Model



Database

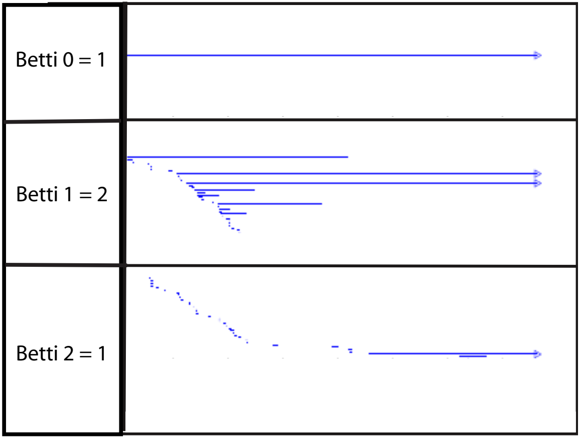


Three Circle Model

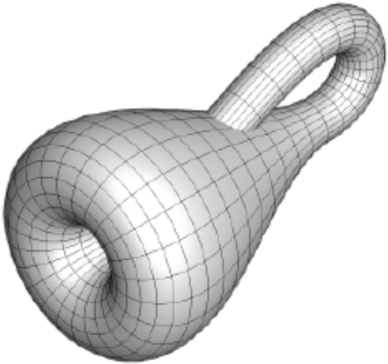
IS THERE A TWO DIMENSIONAL SURFACE IN WHICH THIS PICTURE FITS?

Klein Bottle

4.5×10^6 points, $k = 100$, $T = 10$



Klein Bottle



$\mathcal{K}$ - KLEIN BOTTLE

Klein Bottle

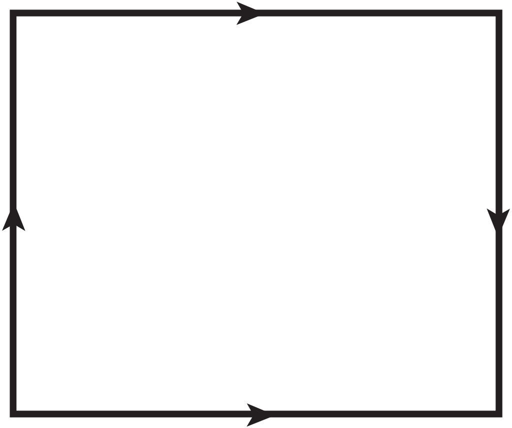
i	0	1	2
$\beta_i(\mathcal{K})$	1	2	1

Klein Bottle

i	0	1	2
$\beta_i(\mathcal{K})$	1	2	1

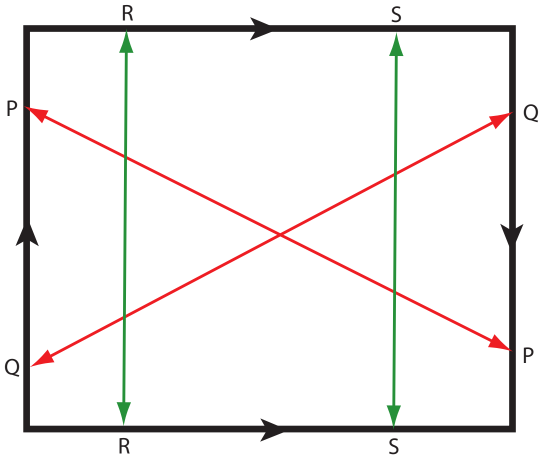
Agrees with the Betti numbers we found from data

Klein Bottle



Identification Space Model

Klein Bottle

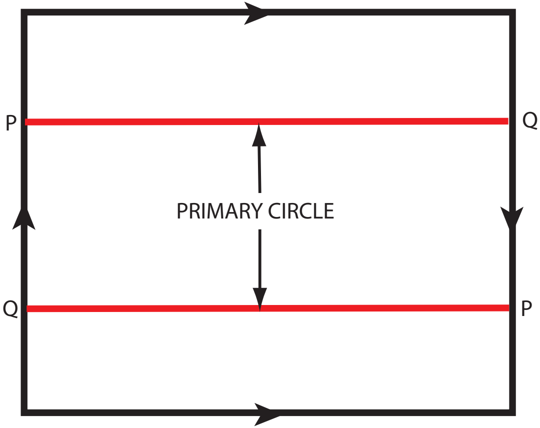


Identification Space Model

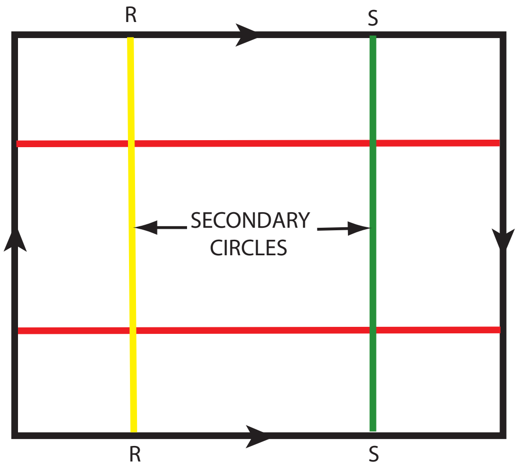
Klein Bottle

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

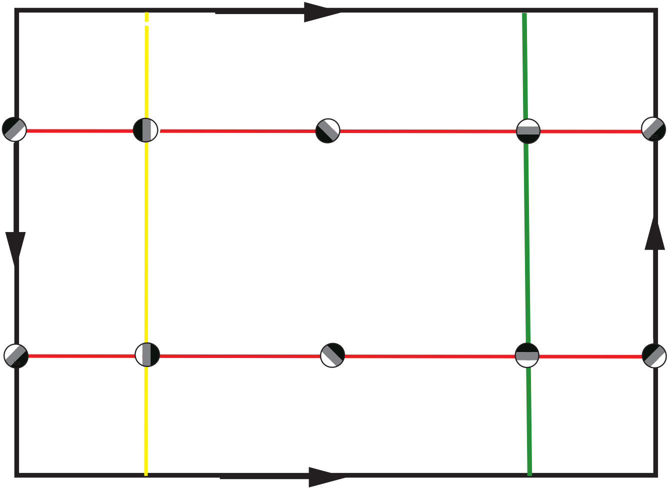
Klein Bottle



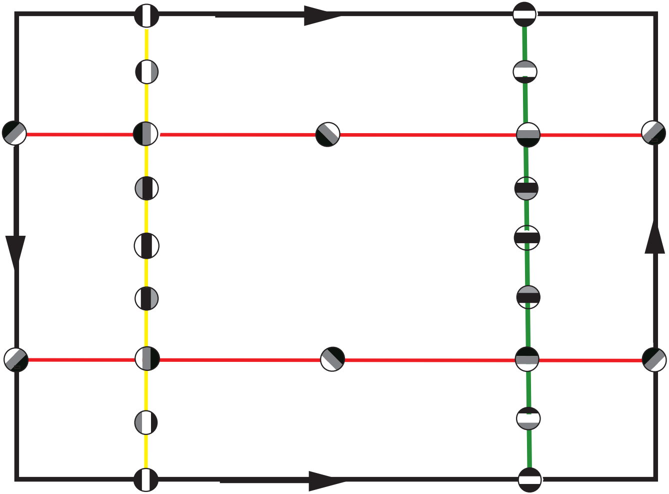
Klein Bottle



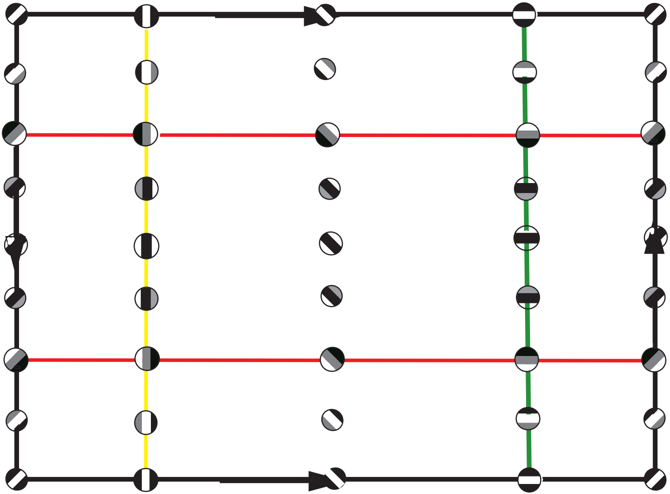
Mapping Patches



Mapping Patches



Mapping Patches



Natural Image Statistics

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. λ is a linear functional
3. $\int_D f = 0$
4. $\int_D f^2 = 1$

Kleinlet Compression

- This understanding of density can be applied to develop compression schemes

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.

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

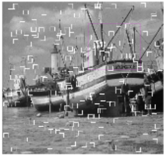
Kleinlet

Wedgelet

Kleinlet Compression

PSNR Comparisons

Kleinlets | Wedges



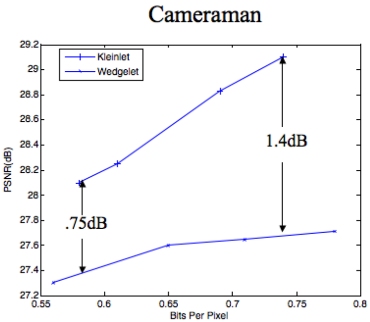
16x16 patches on a 512x512 image

PSNR=24.4

PSNR=22.9

Kleinlet Compression

Compression comparison between kleinlets and wedgelets



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

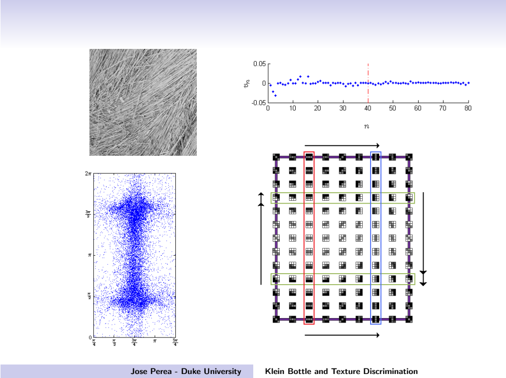
Texture Recognition

- 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



Summary

- Compression and texture recognition often obtained by using finite dictionaries

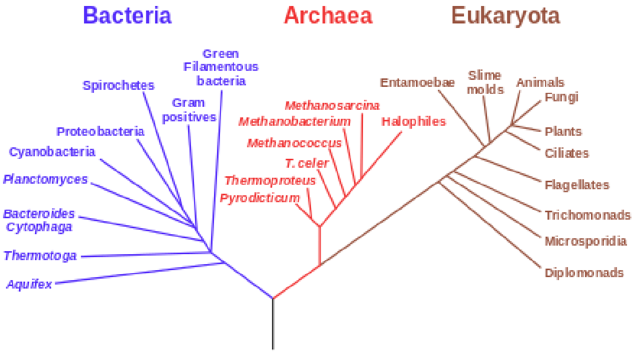
Summary

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

Summary

- 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



Tree of Life

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

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*

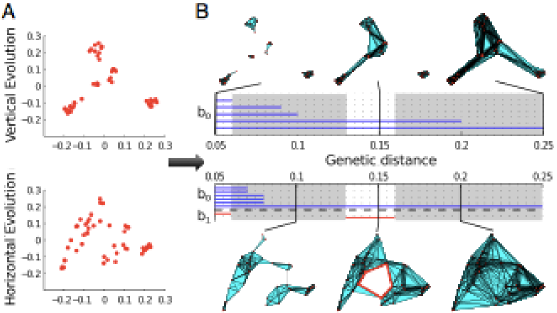
Evolution

- 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 ?

Evolution

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 suggests standard model is incomplete

Evolution

- Can study persistence barcodes of metric spaces of trees arising in evolution
- Presence of large loops can suggests 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 suggests 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 suggests 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
- J. Chan, G. C., and R. Rabadan, Proc. Natl. Acad. Sci. 2013

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 $\mathfrak{B}$ under "bottleneck distance"

Other Applications of Persistence

- Barcodes form a metric space $\mathfrak{B}$ under "bottleneck distance"
- Each barcode $\beta_k(-)$ is Lipschitz with constant one from metric spaces with Gromov-Hausdorff metric to $\mathfrak{B}$

Other Applications of Persistence

- Barcodes form a metric space $\mathfrak{B}$ under "bottleneck distance"
- Each barcode $\beta_k(-)$ is Lipschitz with constant one from metric spaces with Gromov-Hausdorff metric to $\mathfrak{B}$
- $\mathfrak{B}$ is also an infinite algebraic variety, suitably defined

Other Applications of Persistence

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

Other Applications of Persistence

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

Thank you!