Math 149: Applied Algebraic Topology
Introductory Lecture

Gunnar Carlsson, Stanford University

January 7, 2014

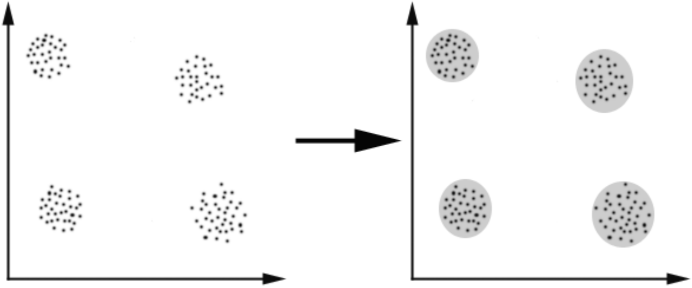
Shape of Data

- Data has shape

Shape of Data

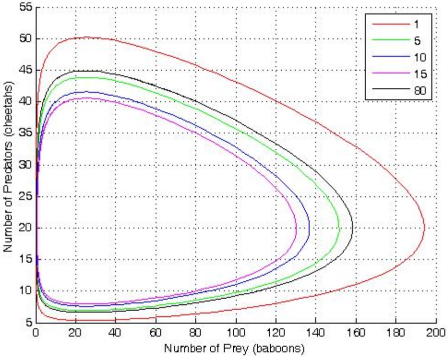
- Data has shape
- The shape matters

Shape of Data



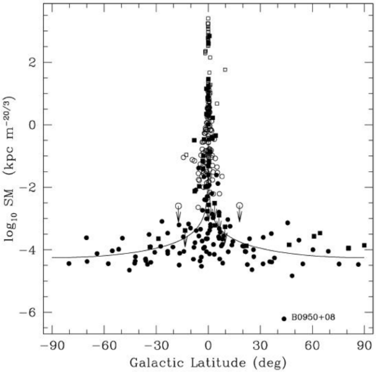
Clusters

Shape of Data

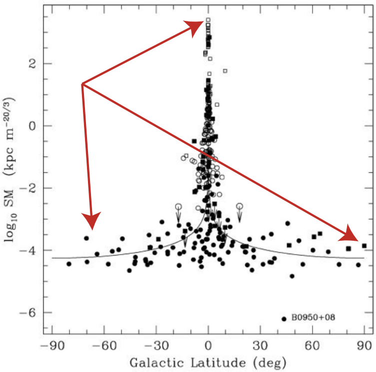


Predator-Prey model

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

Three key ideas:

Topology

Three key ideas:

- Coordinate freeness

Topology

Three key ideas:

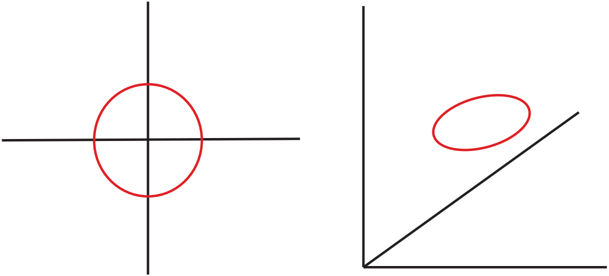
- Coordinate freeness
- Invariance under deformation

Topology

Three key ideas:

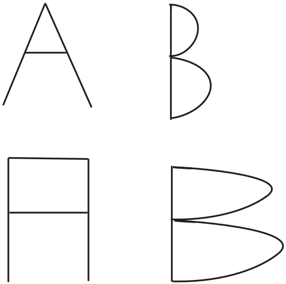
- Coordinate freeness
- Invariance under deformation
- Compressed representations

Topology



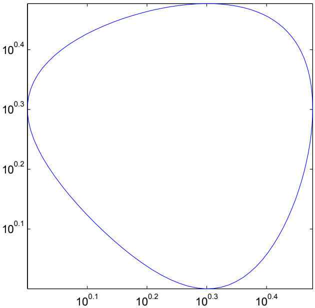
Coordinate Freeness

Topology



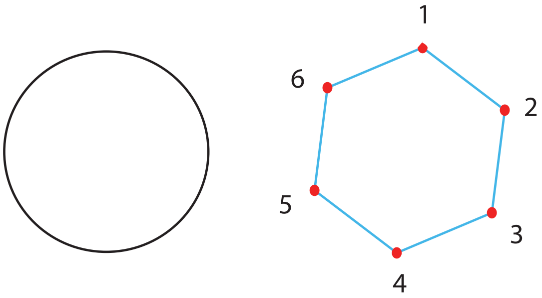
Invariance to Deformations

Topology



Log-log plot of a circle in the plane

Topology



Compressed Representations of Geometry

Topology

Two tasks:

Topology

Two tasks:

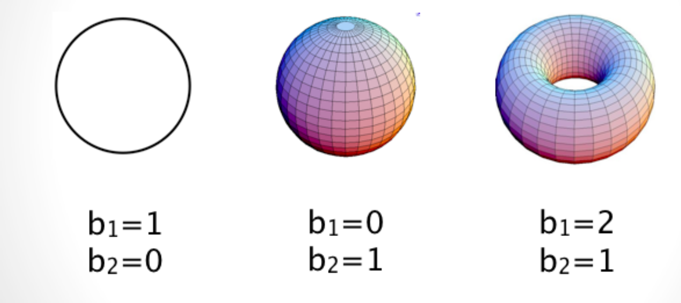
- Measure shape

Topology

Two tasks:

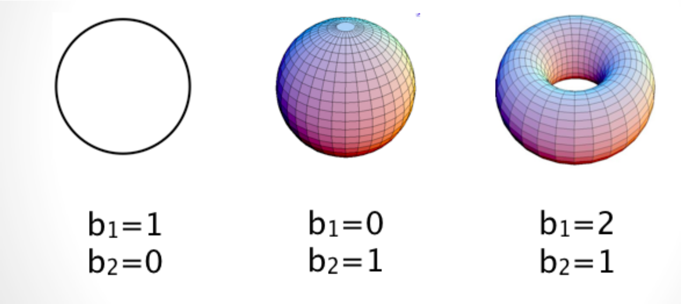
- Measure shape
- Represent shape

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 $f : X \rightarrow Y$ induces linear transformation $H_i(f) : H_i(X) \rightarrow H_i(Y)$

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?

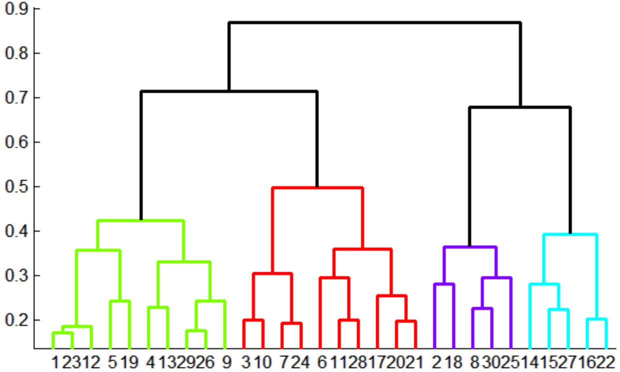
Measuring Shape of Data

- 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 suggest an approach

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

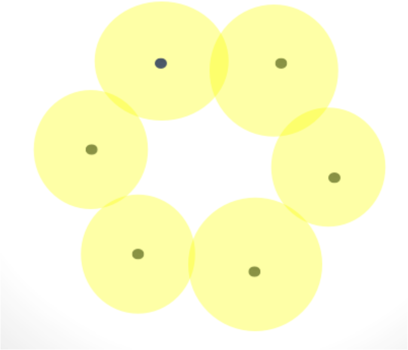
Measuring Shape of Data

- How to build spaces from finite metric spaces

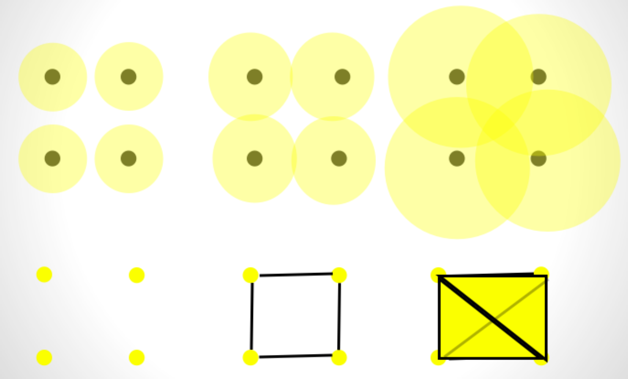
Measuring Shape of Data

- 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

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

Measuring Shape of Data

- Provides an increasing sequence of simplicial complexes
- Apply H_i
- Gives a diagram of vector spaces

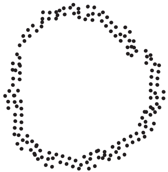
$$V_0 \to V_1 \to V_2 \to V_3 \to \cdots$$

- Call such algebraic structures *persistence vector spaces*

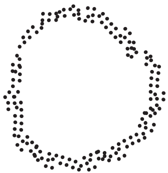
Measuring the Shape of Data

Can we classify persistence vector spaces, up to isomorphism?

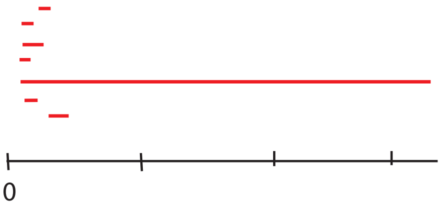
Measuring the Shape of Data - Barcodes



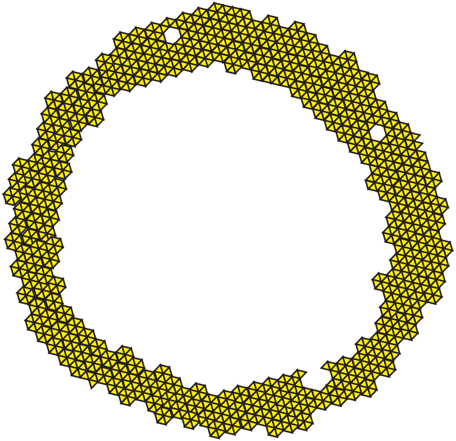
Measuring the Shape of Data - Barcodes



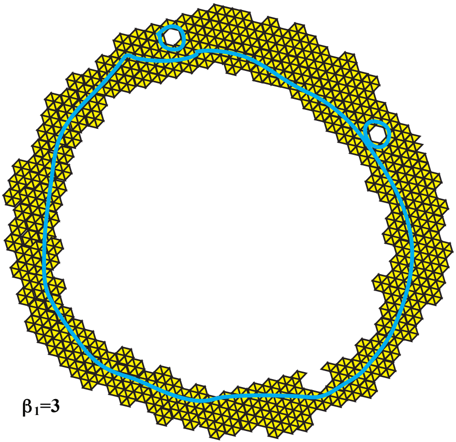
One dimensional barcode:



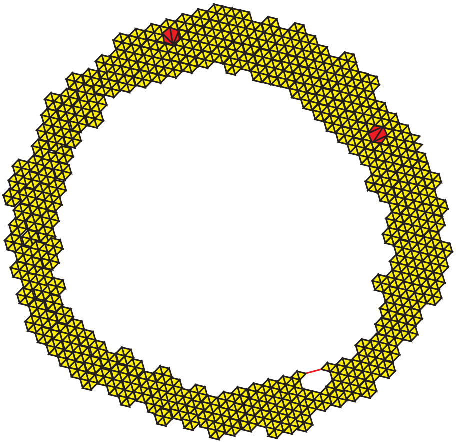
Measuring the Shape of Data - Barcodes



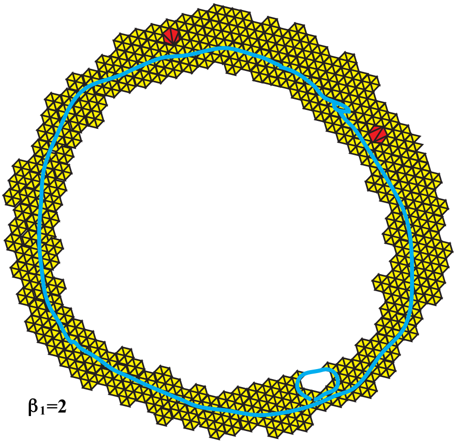
Measuring the Shape of Data - Barcodes



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)

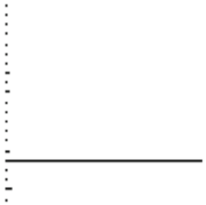
Typical camera uses tens of thousands of pixels, so images lie in a very high dimensional space, call it *pixel space*, $\mathcal{P}$

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?

Primary Circle

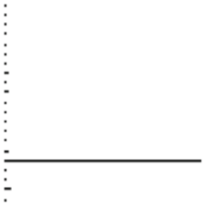
$$5 \times 10^4 \text{ points}, \ k = 300, \ T = 25$$



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

Primary Circle

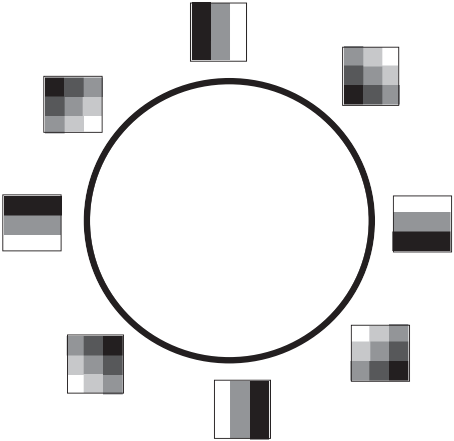
$$5 \times 10^4 \text{ points, } k = 300,\ T = 25$$



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 $\beta_1 = 5$

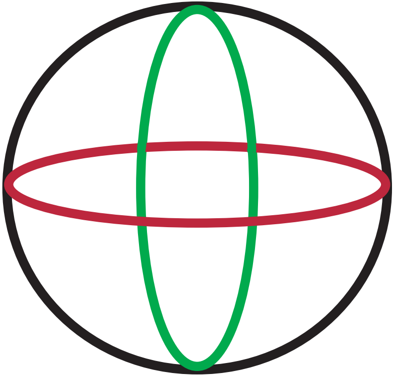
Three Circle Model

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

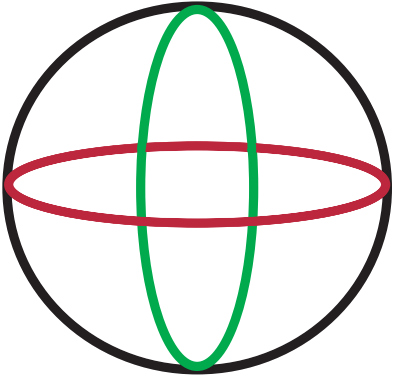
One-dimensional barcode, suggests $\beta_1 = 5$

What's the explanation for this?

Three Circle Model

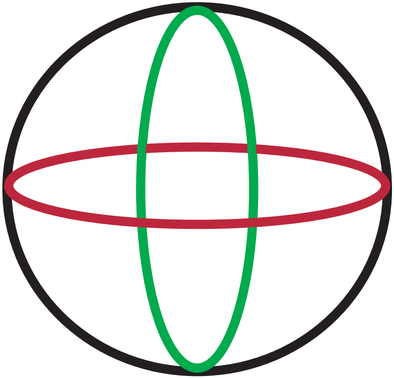


Three Circle Model



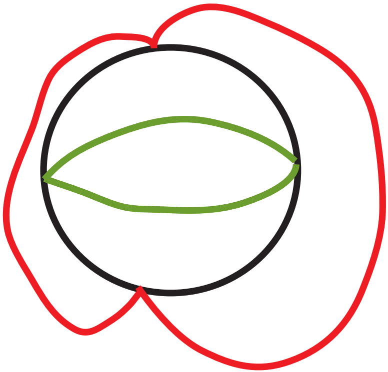
THREE CIRCLE MODEL

Three Circle Model

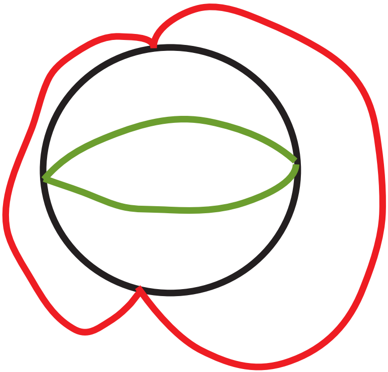


Red and green circles do not touch, each touches black circle

Three Circle Model

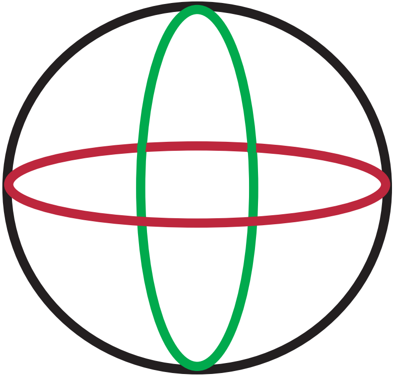


Three Circle Model



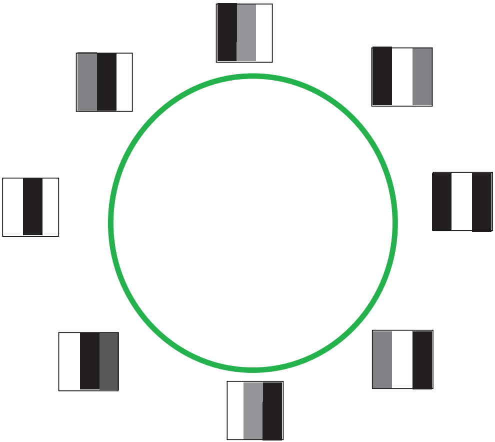
$$\beta_1 = 5$$

Three Circle Model



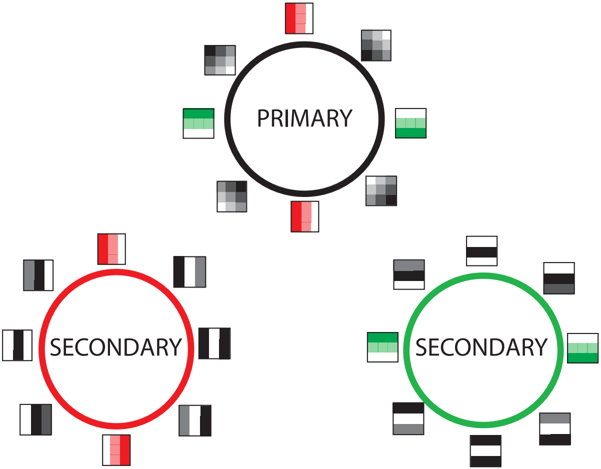
Does the data fit with this model?

Three Circle Model

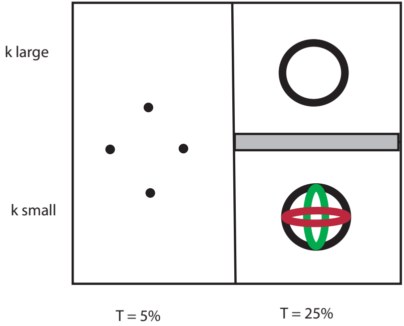


SECONDARY CIRCLE

Three Circle Model



Database

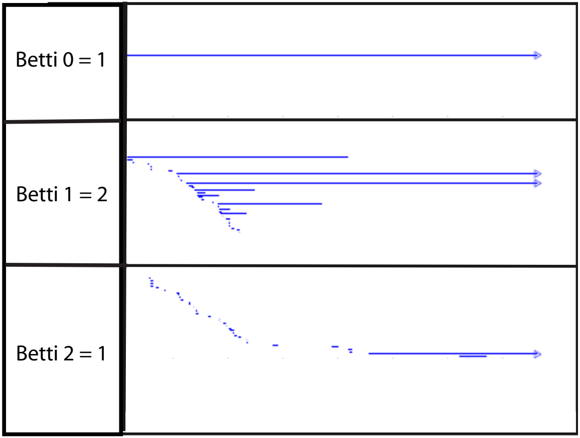


Three Circle Model

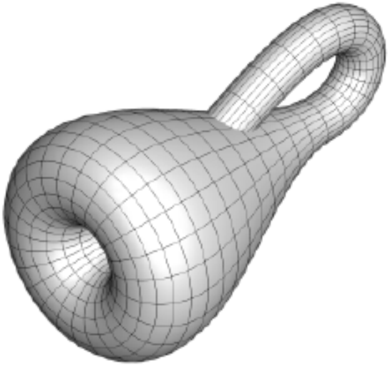
IS THERE A TWO DIMENSIONAL SURFACE IN WHICH THIS PICTURE FITS?

Klein Bottle

4.5×10^6 points, $k = 100$, $T = 10$



Klein Bottle



$\mathcal{K}$ - KLEIN BOTTLE

Klein Bottle

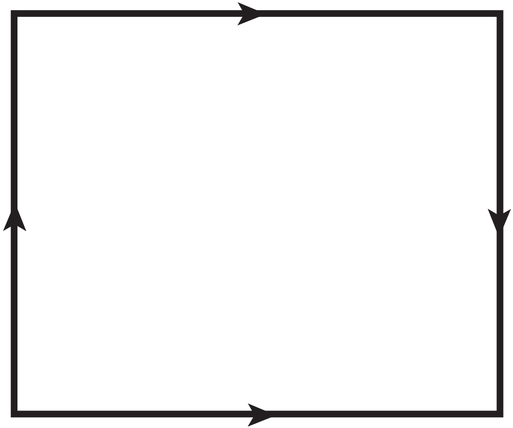
i	0	1	2
$\beta_i(\mathcal{K})$	1	2	1

Klein Bottle

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

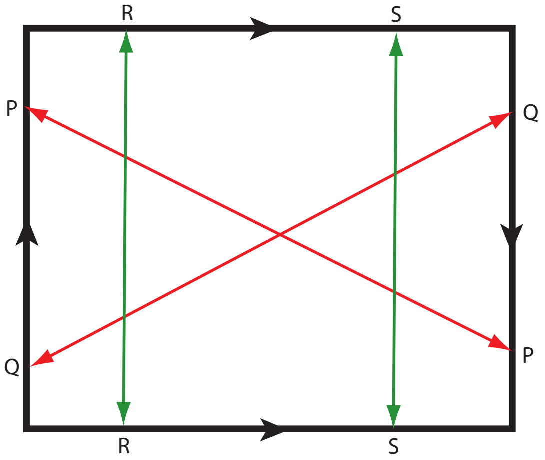
Agrees with the Betti numbers we found from data

Klein Bottle



Identification Space Model

Klein Bottle

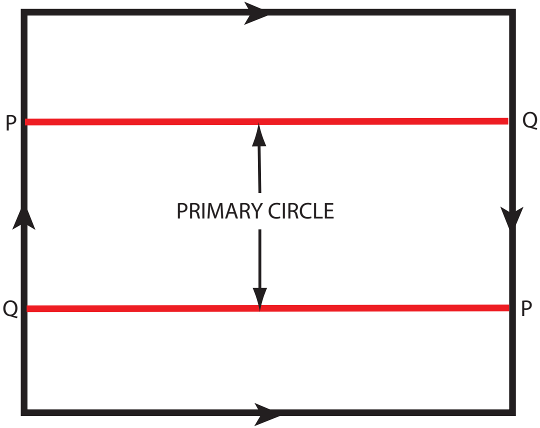


Identification Space Model

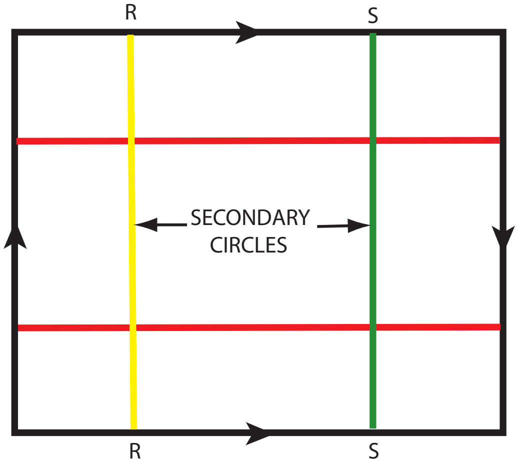
Klein Bottle

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

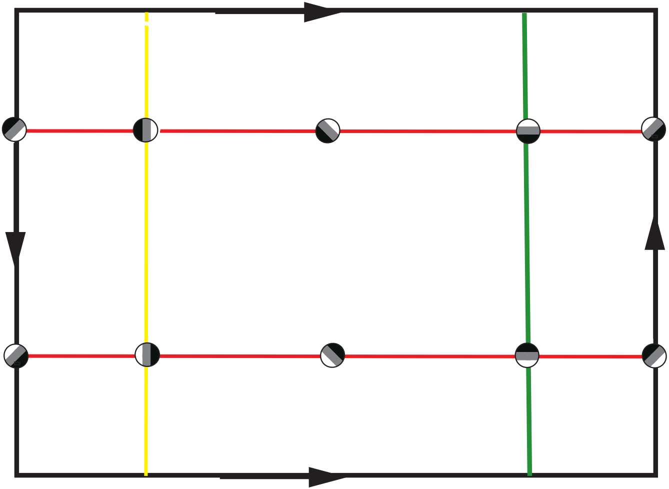
Klein Bottle



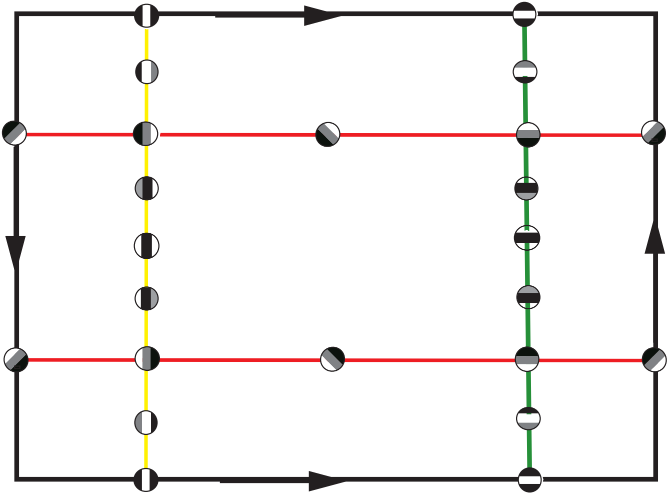
Klein Bottle



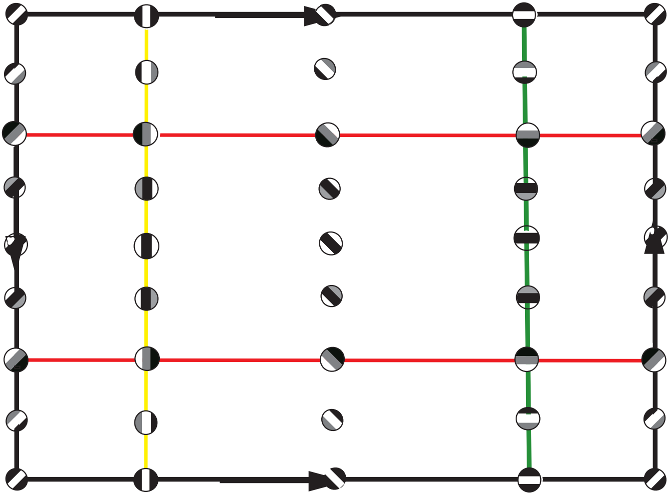
Mapping Patches



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

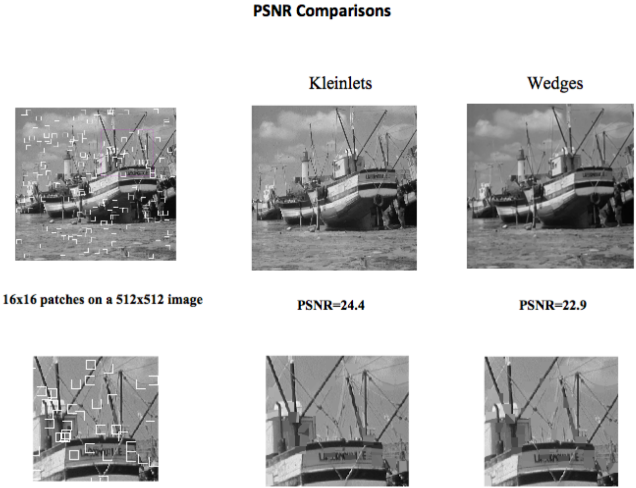
Kleinlet

Wedgelet

Kleinlet

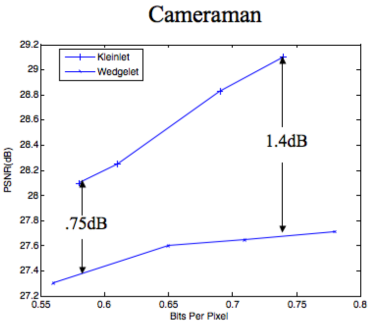
Wedgelet

Kleinlet Compression



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

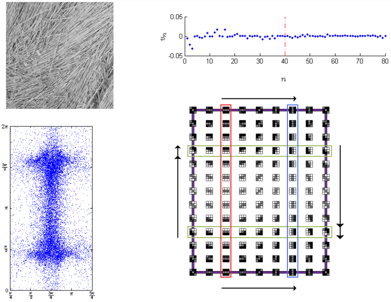
Texture Recognition

- 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



Other Applications of Persistence

- Homology only detects homotopy equivalence

Other Applications of Persistence

- Homology only detects homotopy equivalence
- Can we find persistent methods which capture more about the point cloud?

Other Applications of Persistence

- Homology only detects homotopy equivalence
- Can we find persistent methods which capture more about the point cloud?
- Methods from manifold topology can help

Other Applications of Persistence

- Instead of using a scale parameter, we can use a function on the data set as basis for filtering a Vietoris-Rips complex

Other Applications of Persistence

- Instead of using a scale parameter, we can use a function on the data set as basis for filtering a Vietoris-Rips complex
- Requires that we fix the scale parameter

Other Applications of Persistence

- Instead of using a scale parameter, we can use a function on the data set as basis for filtering a Vietoris-Rips complex
- Requires that we fix the scale parameter
- The quantity on which we filter is usually a geometric quantity

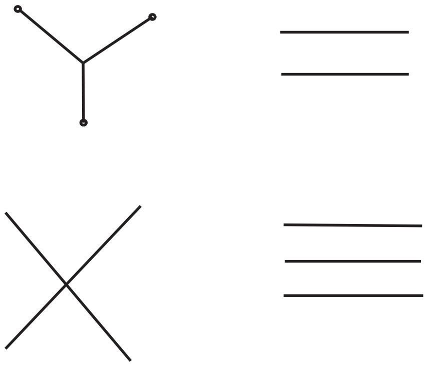
Borel-Moore for Point Clouds

- Point clouds are finite, so doesn't make direct sense
- Replace the ends with some kind of boundary for the space
- Define data depth function on a point cloud X as

$$\Delta(x) = \sum_{x' \in X} d(x, x')$$

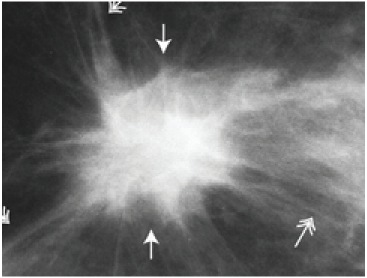
- Define the boundary ∂X as the set of local minima for Δ.
- Borel Moore is now the relative homology of $(X, \partial X)$
- Persistent version: use persistence based on $-\Delta$ instead of scale parameter.

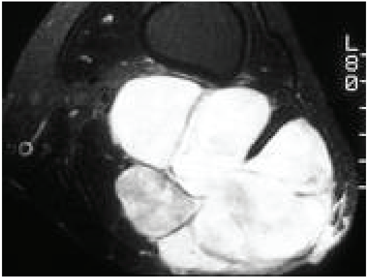
Borel-Moore for Point Clouds



Can now distinguish between "Y" and "X", even though they are homotopy equivalent

Borel-Moore: Shape of Tumors





Spiculated

Lobulated

Sharpening Homology

- General principle: apply homology to (filtered) spaces constructed from the given space using geometric information

Applications of Persistence

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

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

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 $\mathfrak{P}$ from $\mathfrak{M}$ (space of metric spaces with Gromov-Hausdorff metric) to $\mathfrak{B}$ (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 *barcode space*, has a metric
- Persistence gives a map $\mathfrak{P}$ from $\mathfrak{M}$ (space of metric spaces with Gromov-Hausdorff metric) to $\mathfrak{B}$ (barcode space with bottleneck distance)
- $\mathfrak{P}$ is distance non-increasing (Chazal, Mémoli, Guibas, Oudot)

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 $\mathfrak{P}$ from $\mathfrak{M}$ (space of metric spaces with Gromov-Hausdorff metric) to $\mathfrak{B}$ (barcode space with bottleneck distance)
- $\mathfrak{P}$ 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 $\mathfrak{P}$ from $\mathfrak{M}$ (space of metric spaces with Gromov-Hausdorff metric) to $\mathfrak{B}$ (barcode space with bottleneck distance)
- $\mathfrak{P}$ 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.

Representing Shape

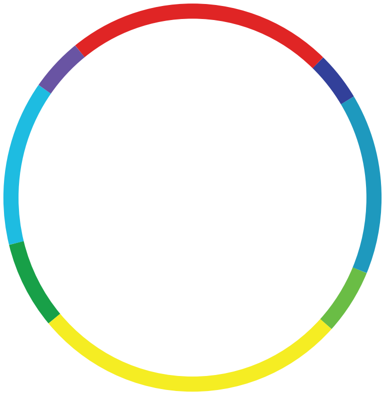
Can one extend topological mapping methods (compressed representations) from idealized shapes to data?

Representing Shape

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

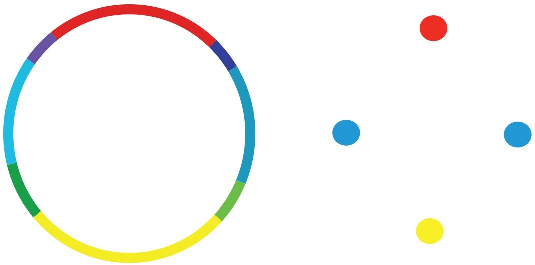
Yes (Singh, Memoli, G. C.)

Topological Mapping



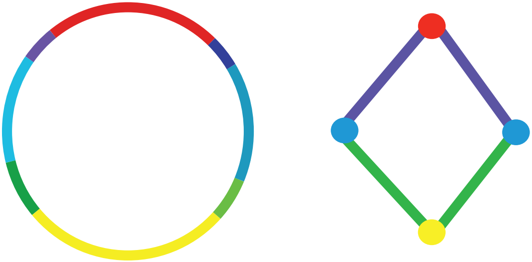
Covering of Circle

Topological Mapping



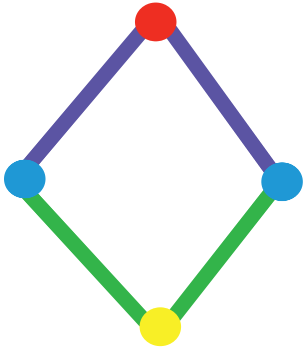
Create nodes

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.

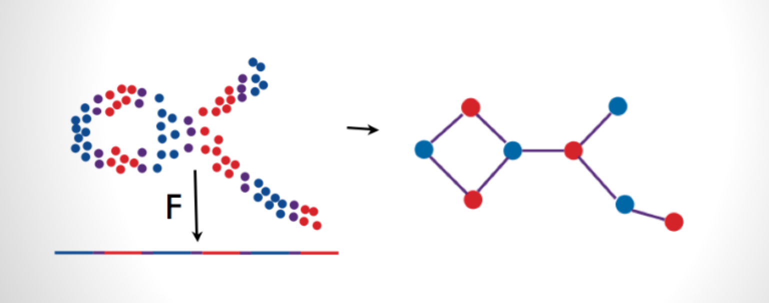
Mapping

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

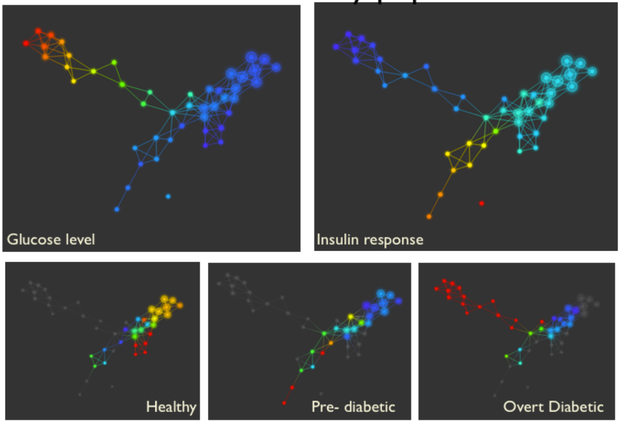
Mapping

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

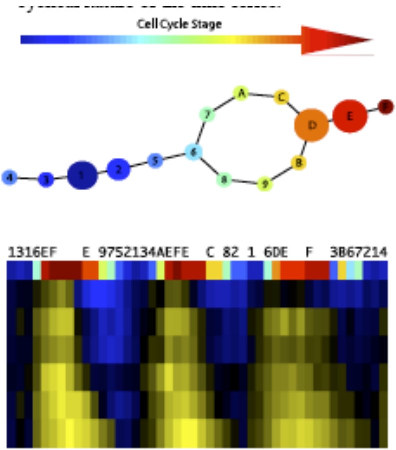
Mapping

Relationships between diabetic, pre-diabetic and healthy populations



Miller-Reaven Diabetes Dataset

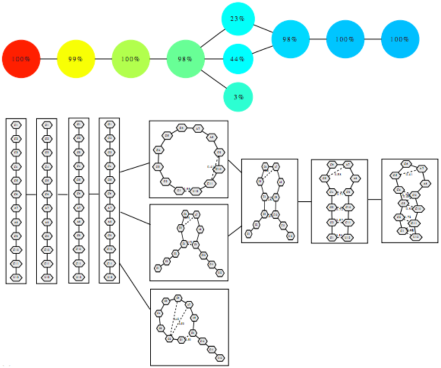
Mapping



Cell Cycle Microarray Data

Joint with M. Nicolau, Nagarajan, G. Singh

Mapping



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

Mapping

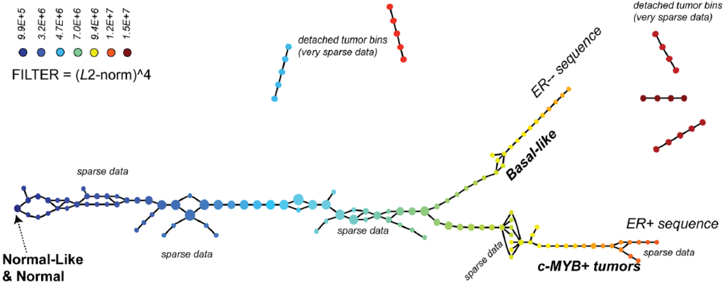
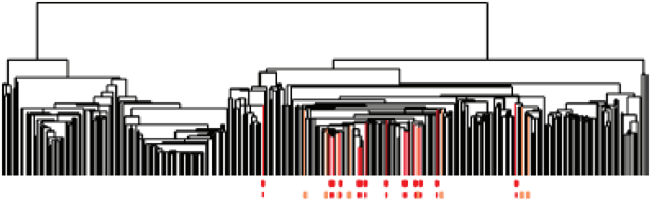
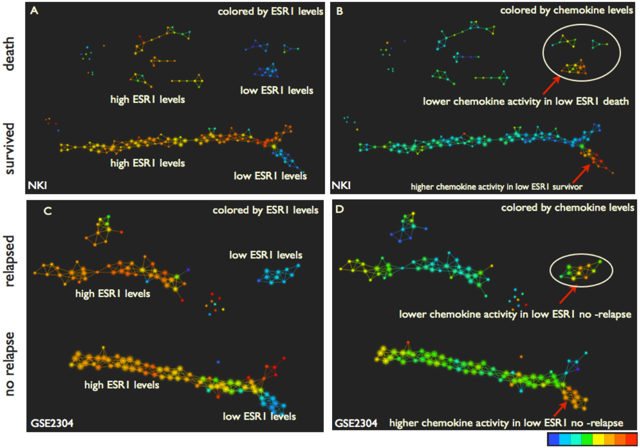


Diagram of gene expression profiles for breast cancer
M. Nicolau, A. Levine, and G. Carlsson, PNAS 2011

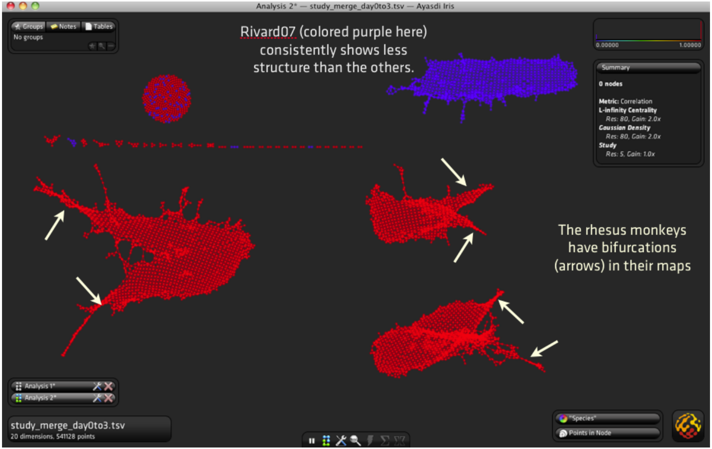
Mapping



Comparison with hierarchical clustering



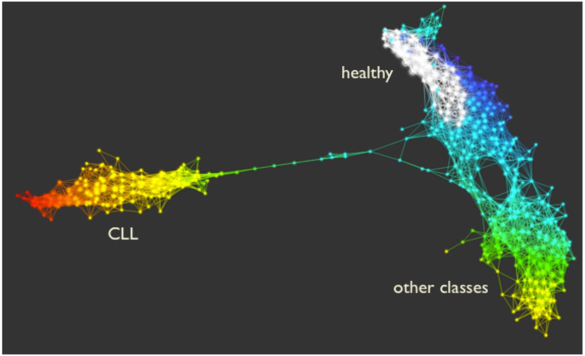
Different platforms - importance of coordinate free approach



Different platforms - importance of coordinate free approach

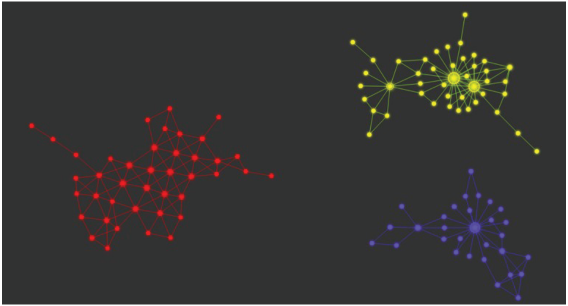
Mapping

Topological structure of leukemia



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

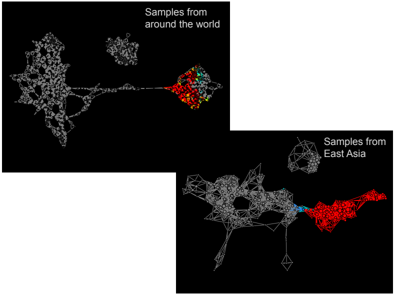
Mapping



Serendipity - copy number variation reveals parent child relations

Mapping

Example: DNA Sequencing



About the Data

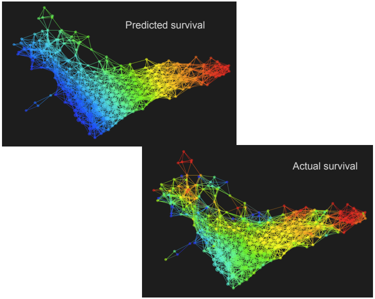
DNA sequencing data includes hundreds of thousands of categorical features, where similarity, or distance, between samples is hard to define. However, this data provides the opportunity to relate genotypes with disease.

1,092 samples
250,000 columns

Using Iris features designed for categorical data, Iris networks map populations from around the world (upper left) and of three subpopulations in East Asia (lower right). Both networks show the distribution of Japanese samples within the network.

Mapping

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.

Thank You!