The Mapper Algorithm

Western TDA Learning Seminar

Andrew W. Herring

Department of Mathematics
Western University

June 7, 2018

Preliminaries

Topological Mapper

Statistical Mapper

Sanity Checks

Applications

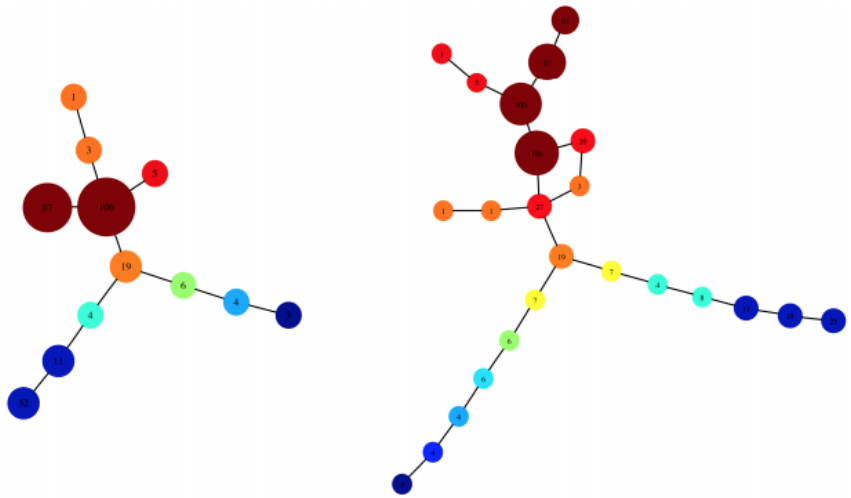
Mapper is an important tool used in TDA for data visualization.

Input

- point cloud;
- "filter function;"
- covering of a metric space;
- clustering algorithm;
- various other parameters.

Output

Graph (or higher simplicial complex) which is thought to capture aspects of the topology of the point cloud.



Developed by Singh, Mémoli, and Carlsson in [6], Mapper gives a multi-resolution, low dimensional picture of the point cloud. It's highly customizable, and has a track record of revealing structure that clustering and (linear or nonlinear) "projection pursuit" methods miss.

Image Source: Singh, Mémoli, Carlsson [6]

Notation

Let $n \geq 1$ be an integer, let $[n] := \{0, \ldots, n\}$.

- An n-**simplex** σ is the convex hull of $n+1$ affinely independent **vertices** $S = \{v^i\}_{i \in [n]}$ in $\mathbb{R}^d$ where $d \geq n$.
- A simplex τ defined by $T \subseteq S$ is called a **face**.
- A **simplicial complex** K is a finite set of simplices which meet along faces, every one of which is in K.
- Let e^0 denote the origin in $\mathbb{R}^n$ and e^i the i-th standard basis vector for $\mathbb{R}^n$.
- The **standard** n-**simplex** $\Delta^n \subset \mathbb{R}^n$ is the convex hull of $\{e^i\}_{i \in [n]}$.
- Given any subset $J \subseteq [n]$, let Δ^J be the face of Δ^n spanned by $\{e^j\}_{j \in J}$.

Nerves

Let X be a topological space and $\mathcal{U} = \{U_\alpha\}_{\alpha \in A}$ a covering of X.
The **nerve of** $\mathcal{U}$, denoted $N\mathcal{U}$, is the abstract simplicial complex with vertex set A and where $\{\alpha_0, \ldots, \alpha_k\}$ spans a k-simplex if and only if

$$U_{\alpha_0} \cap \ldots \cap U_{\alpha_k} \neq \emptyset$$

Nerve Theorem, [2]

X and $\mathcal{U}$ as above, $\mathcal{U}$ a covering by open sets which is enumerable. Suppose further that for all $\emptyset \neq S \subseteq A$ we have that $\bigcap_{s \in S} U_s$ is either empty or contractible. Then $N\mathcal{U}$ is homotopy equivalent to X.

Coverings and complexes

Suppose (X, d) metric space and fix $\varepsilon > 0$. Further suppose that $V \subseteq X$ has the property that

$$X = \bigcup_{v \in V} B_\varepsilon(v)$$

where $B_\varepsilon(v) = \{y \in X : d(v, y) < \varepsilon\}$.
Then $\{B_\varepsilon(v)\}_{v \in V}$ is a covering and has nerve $\check{C}(V, \varepsilon)$, the **Čech complex attached to** (V, ε).

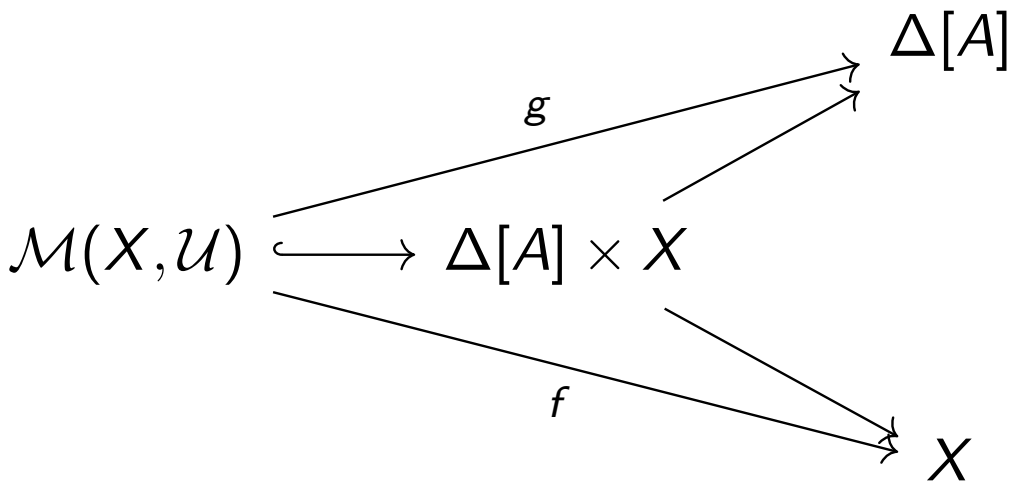
Mayer-Vietoris blowup complex

- X topological space;
- $\mathcal{U} = \{U_\alpha\}_{\alpha \in A}$ a finite covering of X;
- $\Delta[A]$ the standard simplex with vertex set A (i.e. $\Delta[A] = \Delta^{|A|-1}$ and $\dim(\Delta[A]) = |A| - 1$);
- for $\emptyset \neq S \subseteq A$, let $\Delta[S]$ be the face spanned by S;
- $X[S] := \bigcap_{s \in S} U_s \subseteq X$;

The **Mayer-Vietoris blowup complex of X associated to $\mathcal{U}$**, denoted $\mathcal{M}(X, \mathcal{U})$ is

$$\mathcal{M}(X, \mathcal{U}) = \bigcup_{\emptyset \neq S \subseteq A} \Delta[S] \times X[S] \subseteq \Delta[A] \times X$$

Mayer-Vietoris blowup complex



Natural projections satisfy the following properties:

(1) f is a homotopy equivalence when X has the homotopy type of a finite complex and $\mathcal{U}$ is an open covering. Given a partition of unity subordinate to $\mathcal{U}$, can construct an explicit homotopy inverse $\varphi : X \to \mathcal{M}(X,\mathcal{U})$ [7], [5].

(2) g is a homotopy equivalence onto its image ($\check{C}(\mathcal{U})$) when all the $X[S]$ are empty or contractible. (This is how the Nerve Theorem is proved) [2].

A better target

Obtain map

$$X \xrightarrow{\varphi} \mathcal{M}(X,\mathcal{U}) \xrightarrow{g} \check{C}(\mathcal{U})$$

But we can do even better: let $\check{C}^{\pi_0}(\mathcal{U})$ be the nerve of the covering of X by sets which are path connected components of the sets U_α. Thus covering is indexed by pairs (α, ξ) where ξ is a path component of U_α.
Projection $(\alpha, \xi) \mapsto \alpha$ induces a simplicial map

$$\check{C}^{\pi_0}(\mathcal{U}) \to \check{C}(\mathcal{U})$$

and the projections $U_\alpha \mapsto \pi_0(U_\alpha)$ induce a map

$$\mathcal{M}(X,\mathcal{U}) \to \check{C}^{\pi_0}(\mathcal{U})$$

The composite

$$\mathcal{M}(X,\mathcal{U}) \to \check{C}^{\pi_0}(\mathcal{U}) \to \check{C}(\mathcal{U})$$

is exactly the map g from before [2].

Constructing covers

How do we construct a cover of an abstract topological space?
Suppose we're given a (continuous) **reference map** $\rho : X \to Z$, Z some metric space. Given a covering $\mathcal{U}$ of Z, can pull back along ρ and obtain

$$\rho^*\mathcal{U} = \{\rho^{-1}U_\alpha\}_{\alpha \in A}$$

a covering of X.

Example: $\mathcal{U}[R, e]$

$Z = \mathbb{R}$, R, e positive real numbers. $\mathcal{U}[R, e]$ is the collection of all intervals $[kR - e, (k+1)R + e]$, $k \in \mathbb{Z}$. When $e < R/2$, it has covering dimension 1: i.e. there are no non-trivial triple intersections.

Clustering

For X a point cloud (a finite metric space), path connected components are uninteresting.
Analogue of "connected component" is "cluster." We run a clustering algorithm in order to partition a point cloud into clusters: heuristically, "points within a cluster are similar and points from distinct clusters are dissimilar."
Single-linkage clustering algorithm: (according to [2]) Let $\epsilon > 0$ and define blocks of a partition (clusters) of the point cloud as the set of equivalence classes under $\sim_\epsilon$ where

$$x \sim_\epsilon x' \iff d(x, x') \leq \epsilon$$

Notice that the set of clusters is exactly $\pi_0(VR(X, \epsilon))$, the set of connected components of the Vietoris-Rips complex on X at ϵ.

Discrete $\check{C}^{\pi_0}(\mathcal{U})$

(1) Given a point cloud X; a metric space Z; and a reference map $f : X \to Z$.

(2) Select a covering $\mathcal{U}$ of Z. (Eg. $\mathcal{U} = \mathcal{U}[R, e]$ if $Z = \mathbb{R}$).

(3) If $\mathcal{U} = \{U_\alpha\}_{\alpha \in A}$, pull back covering by the sets $X_\alpha = f^{-1}(U_\alpha)$.

(4) Fix $\epsilon > 0$, construct the set of clusters by the single linkage procedure with ϵ on each set X_α. Obtain a covering of X indexed by pairs (α, c) where $\alpha \in A$, and c is one of the clusters of X_α.

(5) Construct simplicial complex $\check{C}^{\pi_0}(\mathcal{U})$ with vertex set the collection of all pairs (α, c), and where $\{(\alpha_0, c_0), (\alpha_1, c_1), \ldots, (\alpha_k, c_k)\}$ spans a k-simplex if and only if the clusters $c_0, \ldots, c_k$ have a common point.

Discrete $\check{C}^{\pi_0}(\mathcal{U})$

Remarks

- Any clustering algorithm can be used in (4).
- If $\mathcal{U}$ has covering dimension $\leq d$ (i.e. for any family $\{\alpha_0, \ldots, \alpha_t\}$ of distinct indices with $t > d$, then $U_{\alpha_0} \cap \ldots \cap U_{\alpha_t} = \emptyset$), then $\check{C}^{\pi_0}(\mathcal{U})$ also has dimension $\leq d$.
- Choosing the parameter $\epsilon > 0$ in (4) is difficult. One can imagine contexts where ϵ should vary with α. Because clusters under single linkage are exactly connected components of $VR(X, \epsilon)$, can actually use barcodes and β_0 to analyze behavior of clusters over varying choices of ϵ and employ the "persistence philosophy". See [2] Section 3.4 for the discussion of "Scale Space."

Multi-resolution structure

$$\mathcal{U} = \{U_\alpha\}_{\alpha \in A},\ \mathcal{V} = \{V_\beta\}_{\beta \in B}$$

two coverings of a topological space Z. A **map of coverings** $\mathcal{U} \to \mathcal{V}$ is given by a set map $\theta : A \to B$ such that

$$U_\alpha \subseteq V_{\theta(\alpha)}$$

for all $\alpha \in A$.

Examples:

(1) The identity set map on $\mathbb{Z}$ yields a map of coverings $\mathcal{U}[R, e] \to \mathcal{U}[R, e']$ whenever $e \leq e'$.

(2) The map of integers $k \mapsto \lfloor k/2 \rfloor$ defines a map of coverings $\mathcal{U}[R, e] \to \mathcal{U}[2R, e]$. It's two to one in the sense that each interval in $\mathcal{U}[2R, e]$ contains two intervals from $\mathcal{U}[R, e]$.

Multi-resolution structure

A clustering algorithm is **functorial** if given an inclusion of point clouds (distance preserving injective set map) $X \hookrightarrow Y$, then the image of each cluster in X is included in one of the clusters in Y. Get induced set map

$$\{\text{clusters of } X\} \to \{\text{clusters of } Y\}$$

If clustering algorithm in (4) is functorial, we'll see that refining a covering (i.e. applying a map of coverings) refines our simplicial complex.

Multi-resolution structure

Suppose we're given:

- a point cloud X;
- $\rho : X \to Z$ a reference map to a metric space Z;
- Two coverings $\mathcal{U} = \{U_\alpha\}_{\alpha \in A}$ and $\mathcal{V} = \{V_\beta\}_{\beta \in B}$ of Z;
- a map of coverings $\theta : \mathcal{U} \to \mathcal{V}$.

By definition of the covering map,

$$U_\alpha \subseteq V_{\theta(\alpha)}$$

for all $\alpha \in A$ and therefore

$$\rho^{-1}(U_\alpha) \subseteq \rho^{-1}(V_{\theta(\alpha)})$$

for all $\alpha \in A$.

Multi-resolution structure

Applying a functorial clustering algorithm to each set $\rho^{-1}(U_\alpha)$ and $\rho^{-1}(V_\beta)$, obtain map of sets from the set of clusters in $\rho^{-1}(U_\alpha)$ to the set of clusters in $\rho^{-1}(V_{\theta(\alpha)})$, hence a map from the vertex set of $\mathcal{M}(X,\mathcal{U})$ to the vertex set of $\mathcal{M}(X,\mathcal{V})$. In fact, this induces a simplicial map

$$\Theta : \mathcal{M}(X,\mathcal{U}) \to \mathcal{M}(X,\mathcal{V})$$

Eg: obtain diagram of simplicial complexes

$$\ldots \to \mathcal{M}(X,\mathcal{U}[R/4,e]) \to \mathcal{M}(X,\mathcal{U}[R/2,e]) \to \mathcal{M}(X,\mathcal{U}[R,e])$$

Moving further to the left, the coverings give a more refined picture of $\mathbb{R}$: the simplicial complex mirrors this refinement of resolution. **Persistence philosophy!**

Mapper implemented

- **Given:** X point cloud, $|X| = N$, filter function $f : X \to \mathbb{R}$.
- Assume we can always compute inter point distances.
- Let I denote the "range" of f: explicitly $I = [m, M] \subset \mathbb{R}$ where

$$m = \min_{x \in X} f(x), \; M = \max_{x \in X} f(x)$$

- Divide I into a set S of smaller intervals (of uniform length) which overlap. Obtain two resolution controlling parameters: ℓ the length of the intervals, and p the percentage overlap between successive intervals.

Mapper implemented

(1) For each interval $I_j \in S$, let

$$X_j := \{x : f(x) \in I_j\}$$

Then the collection of all such X_j is a covering of X.

(2) For each X_j, perform a clustering algorithm to obtain clusters $\{X_{jk}\}$.

(3) Each cluster defines a vertex of our simplicial complex: draw an edge between vertices whenever $X_{jk} \cap X_{lm} \neq \emptyset$.

Example: noisy circle [6]

- The point cloud data is sampled from a noisy circle in $\mathbb{R}^2$
- filter function is $f(x) = \|x - p\|_2$ where p is the left most point in the data set.
- range of f is covered by 5 intervals, $\ell = 1$, $p = 0.20$.
- Nodes are colored by the average value of the filter function over the cluster.

Example: noisy circle [6]

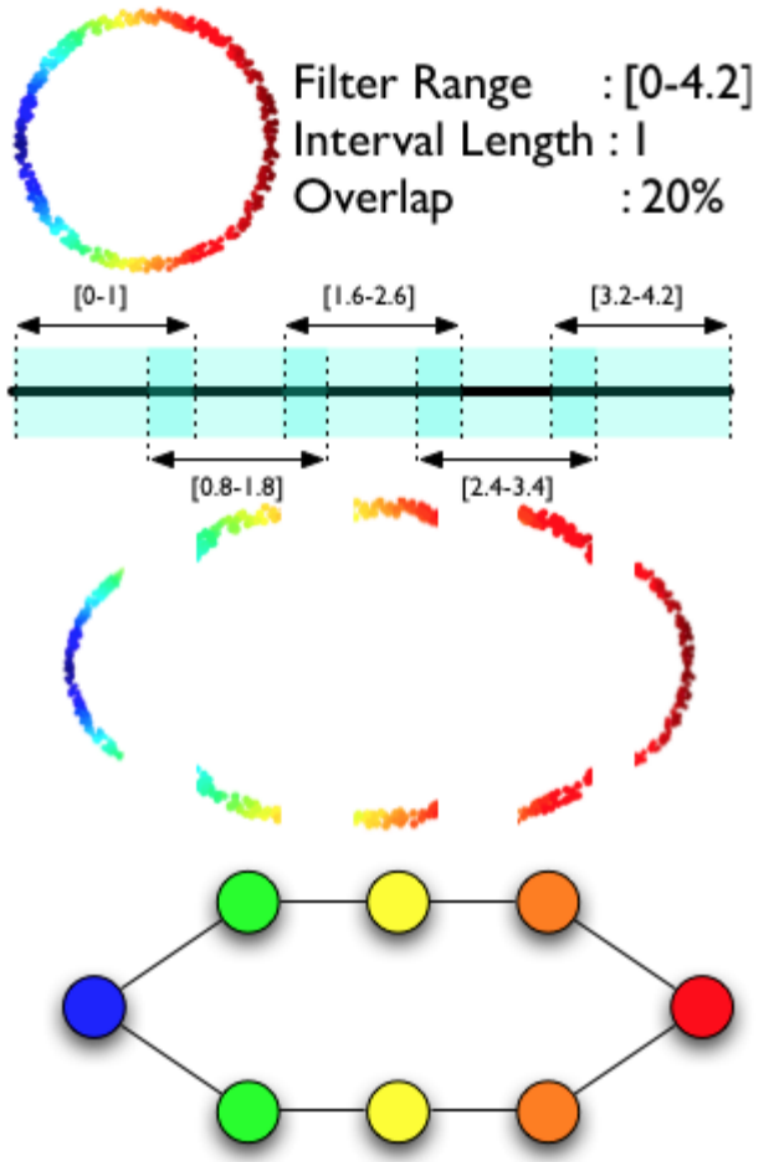


Image Source: Mémoli,Singh,Carlsson [6]

Clustering

Single linkage: $x \sim_\epsilon x'$ iff $d(x, x') \leq \epsilon$, clusters are equivalence classes.
How to choose ϵ?
Heuristically: inter-point distances within clusters should be small, with longer edge lengths required to merge clusters.
Consider $C \in \mathbb{R}^{N-1}$, vector of edge lengths at which the number of clusters successively decreases by one, and histogram of edge lengths in C. Shorter lengths (inter-cluster distances) are distributed smoothly, but edge lengths required to merge natural clusters are disjoint.

Clustering

- Suppose histogram for C has k intervals. Then we find a set of empty intervals after which lengths required to merge clusters appear. Let ϵ be the last length before empty intervals appear, and apply single linkage with that ϵ.
- Increase k to increase number of clusters and vice versa.
- Main limitation is density issues: when clusters with different densities are present, only high density clusters are selected.

Filter functions: density

For $\varepsilon > 0$ estimate the density using a Gaussian kernel

$$f_\varepsilon(x) = C_\varepsilon \sum_{y \in X} \exp\left(\frac{-d(x,y)^2}{\varepsilon}\right)$$

where $x, y \in X$ and C_ε is a constant so that $\int f_\varepsilon(x)\, dx = 1$.
"ε controls the smoothness."

Filter functions: eccentricity

Aim to identify points which are far from some intuitive/natural "center" without having identified one. For $1 \leq p < \infty$

$$E_p(x) = \left(\frac{\sum_{y \in X} d(x,y)^p}{N} \right)^{1/p}$$

and

$$E_\infty(x) = \max_{x' \in X} d(x,x')$$

Then E takes larger values on points further away from "center."

Filter functions: graph Laplacians

Construct a graph from the point cloud: it has vertex set X and edge weights given by $w(x, y) = kd(x, y)$, for k some "smoothing kernel." A **(normalized) graph Laplacian** is

$$L(x, y) = \frac{w(x, y)}{\sqrt{\sum_{z \in X} w(x, z)} \sqrt{\sum_{z \in X} w(y, z)}}$$

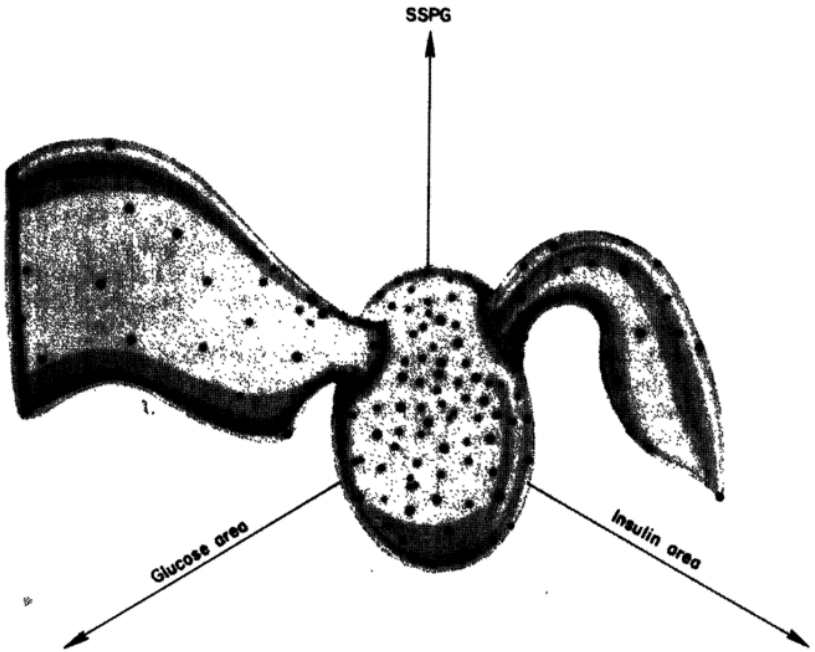
Form the graph laplacian matrix

$$[L(x, y)]_{x,y \in X}$$

and compute its eigenvectors. They're orthogonal, and encode interesting geometric features [6]: use them as filter functions!
(Discrete analogue of Laplace-Beltrami operator on a compact Riemanninan manifold...[1])

Miller-Reaven diabetes study

Applied projection methods to a data set of 145 patients with 6 clinical measurements each. Created a point cloud in $\mathbb{R}^6$ with 145 points, project into $\mathbb{R}^3$ to obtain resulting plot.



Two "flares" represent different diseases: adult and juvenile onset diabetes.

Image Source: Singh, Mémoli, Carlsson [6]

Miller-Reaven diabetes study

Density kernel estimator. On left, three intervals 50% overlap. On right, five intervals 50% overlap. Nodes colored by average density from blue to red. Size of node indicates cluster size. Note the flares!

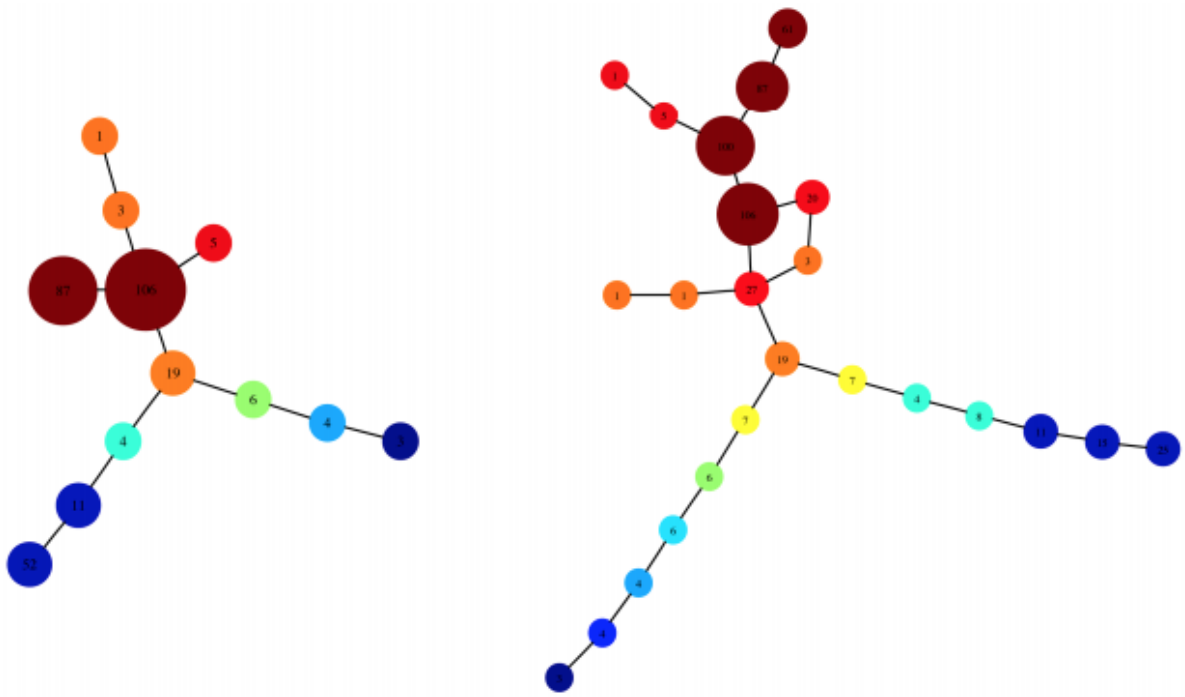


Image Source: Singh, Mémoli, Carlsson [6]

Torus

Generate 1500 evenly sampled points on torus in $\mathbb{R}^3$. Embed in $\mathbb{R}^{30}$ by padding dimensions 4 to 30 with zeros, then apply random rotation to resulting point cloud. Compute first two non-trivial eigenfunctions f_1 and f_2 of Laplacian to use as filter functions (note reference space $\mathbb{R}^2$). Eight intervals each in range of f_i with 50% overlap. Result is cluster of 325 points with 4D simplicial complex. MDS + magic to generate 3D picture (1 and 2D simplices infered from higher simplices.)

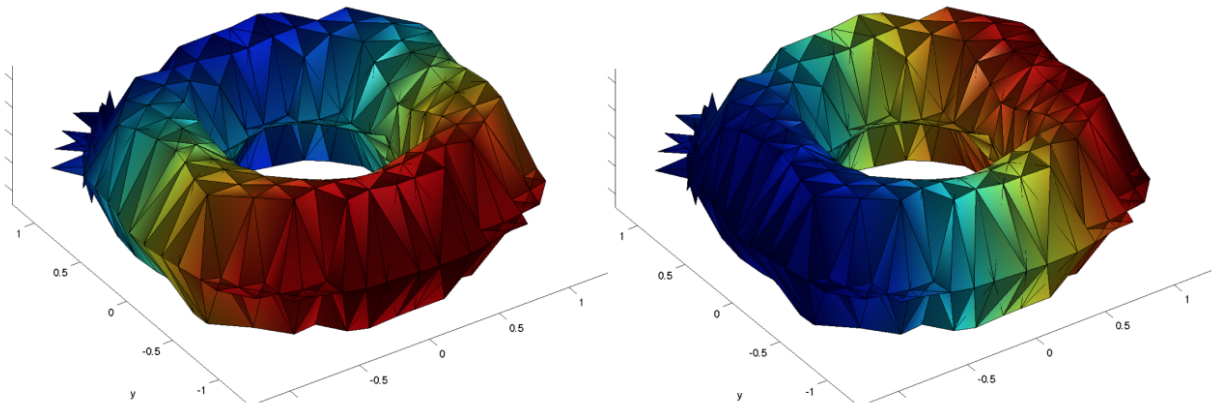


Image Source: Singh, Mémoli, Carlsson [6]

Breast cancer

Metric: correlation distance between gene expression vectors. Filter functions: L^∞ eccentricity and survival variable. Note that the two rows are **DIFFERENT DATA SETS**

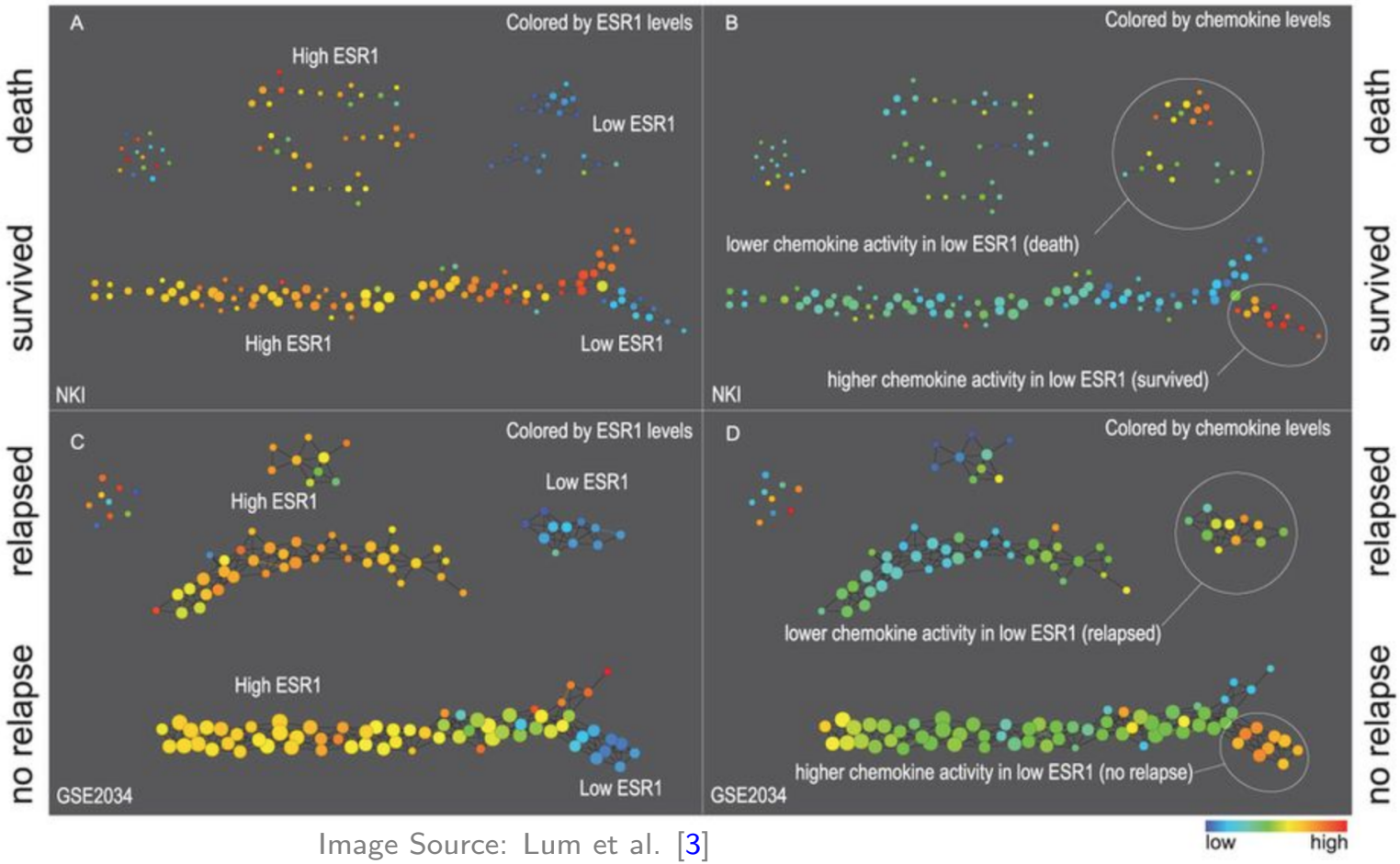


Image Source: Lum et al. [3]

Breast cancer

Mapper finds features that clustering alone misses.

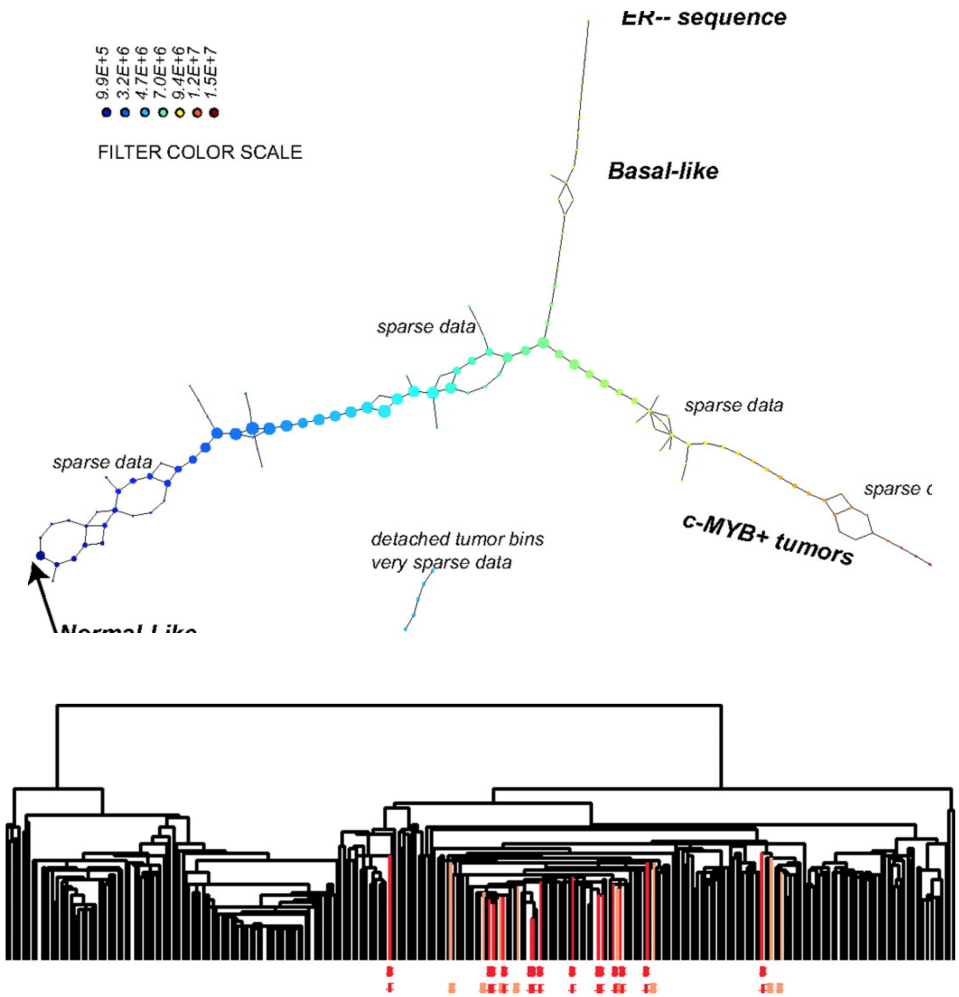


Image Source: Nicolau et al. [4]

NBA

There are more than the traditional five positions in basketball.

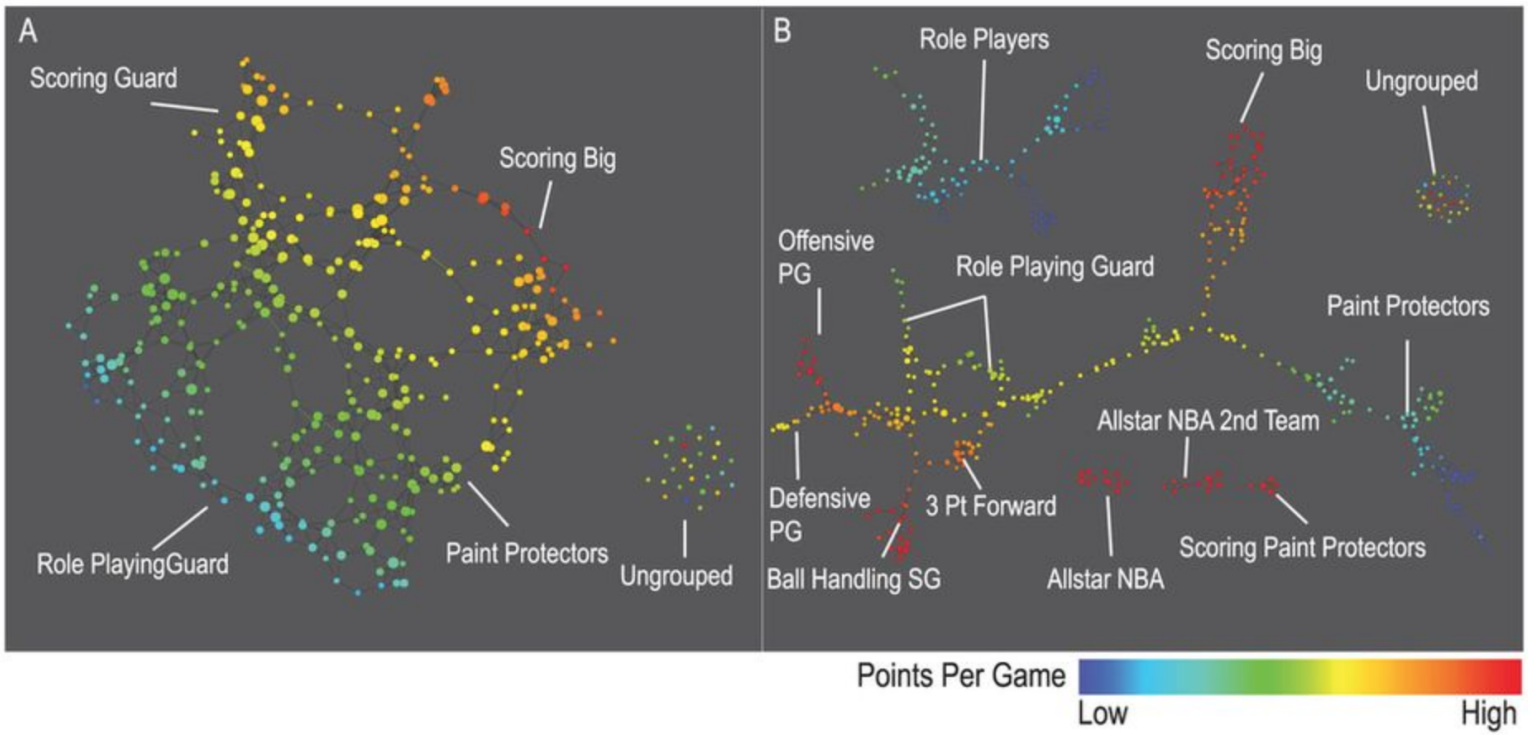


Image Source: Lum et al. [3]

Shape comparison

Triangulated mesh of different poses of the same subjects: camels cats, elephants, faces, heads, horses, and lions. Sampled "landmark points" judiciously. Filter functions E_1 eccentricity. In a statistically meaningful way, mapper outputs from same subject different pose are similar and mapper outputs between different subjects are dissimilar.

Any questions?

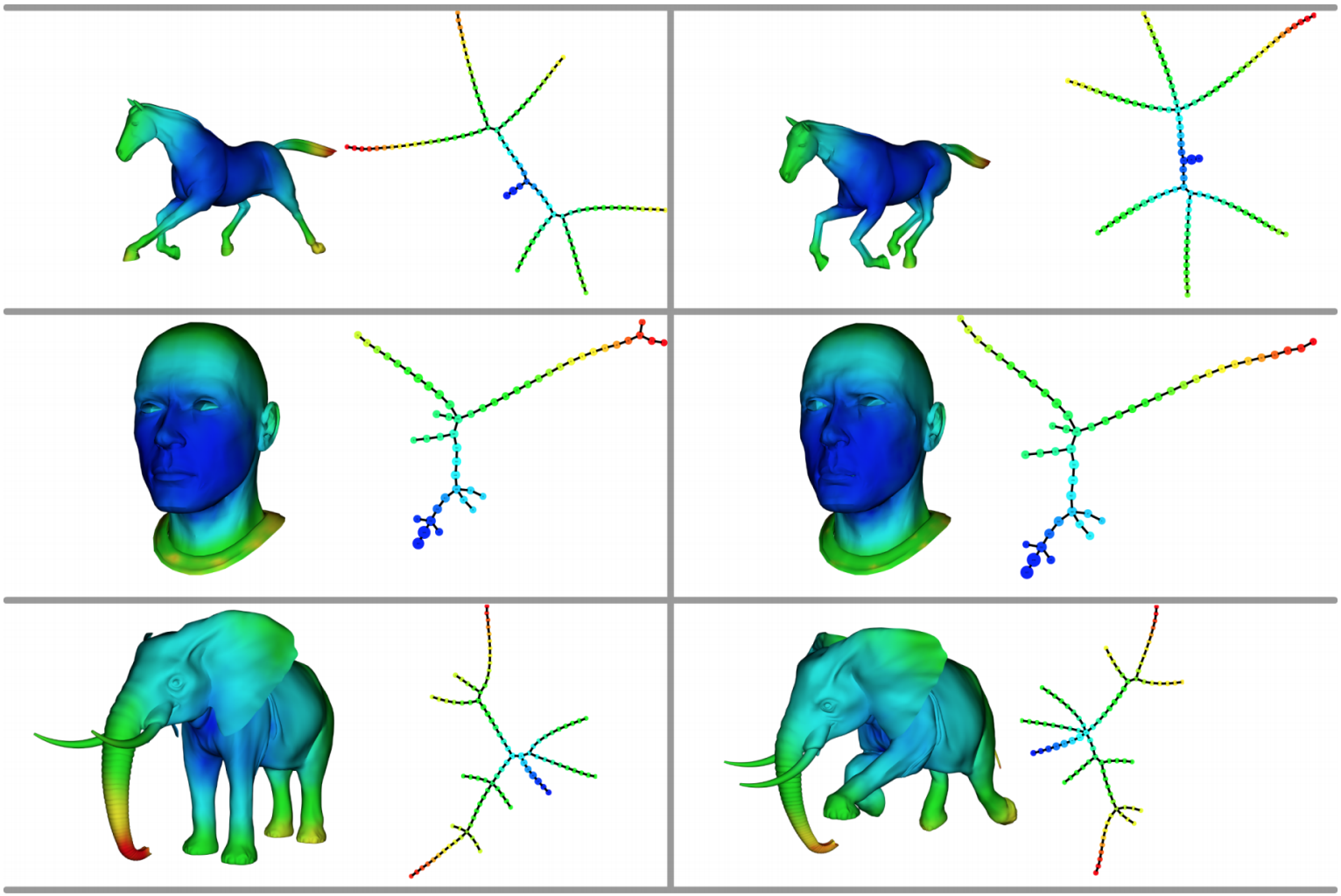


Image Source: Singh, Mémoli, Carlsson [6]

References I

[1] Mikhail Belkin and Partha Niyogi.
Laplacian eigenmaps and spectral techniques for embedding and clustering.
In *Advances in neural information processing systems*, pages 585–591, 2002.

[2] Gunnar Carlsson.
Topology and data.
Bulletin of the American Mathematical Society, 46(2):255–308, 2009.

[3] Pek Y Lum, Gurjeet Singh, Alan Lehman, Tigran Ishkanov, Mikael Vejdemo-Johansson, Muthu Alagappan, John Carlsson, and Gunnar Carlsson.
Extracting insights from the shape of complex data using topology.
Scientific reports, 3:srep01236, 2013.

[4] Monica Nicolau, Arnold J Levine, and Gunnar Carlsson.
Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival.
Proceedings of the National Academy of Sciences, 108(17):7265–7270, 2011.

[5] Graeme Segal.
Classifying spaces and spectral sequences.
Publications Mathématiques de l'IHÉS, 34:105–112, 1968.

[6] Gurjeet Singh, Facundo Mémoli, and Gunnar E Carlsson.
Topological methods for the analysis of high dimensional data sets and 3d object recognition.
In *SPBG*, pages 91–100, 2007.

[7] Afra Zomorodian and Gunnar Carlsson.
Localized homology.
Computational Geometry, 41(3):126–148, 2008.