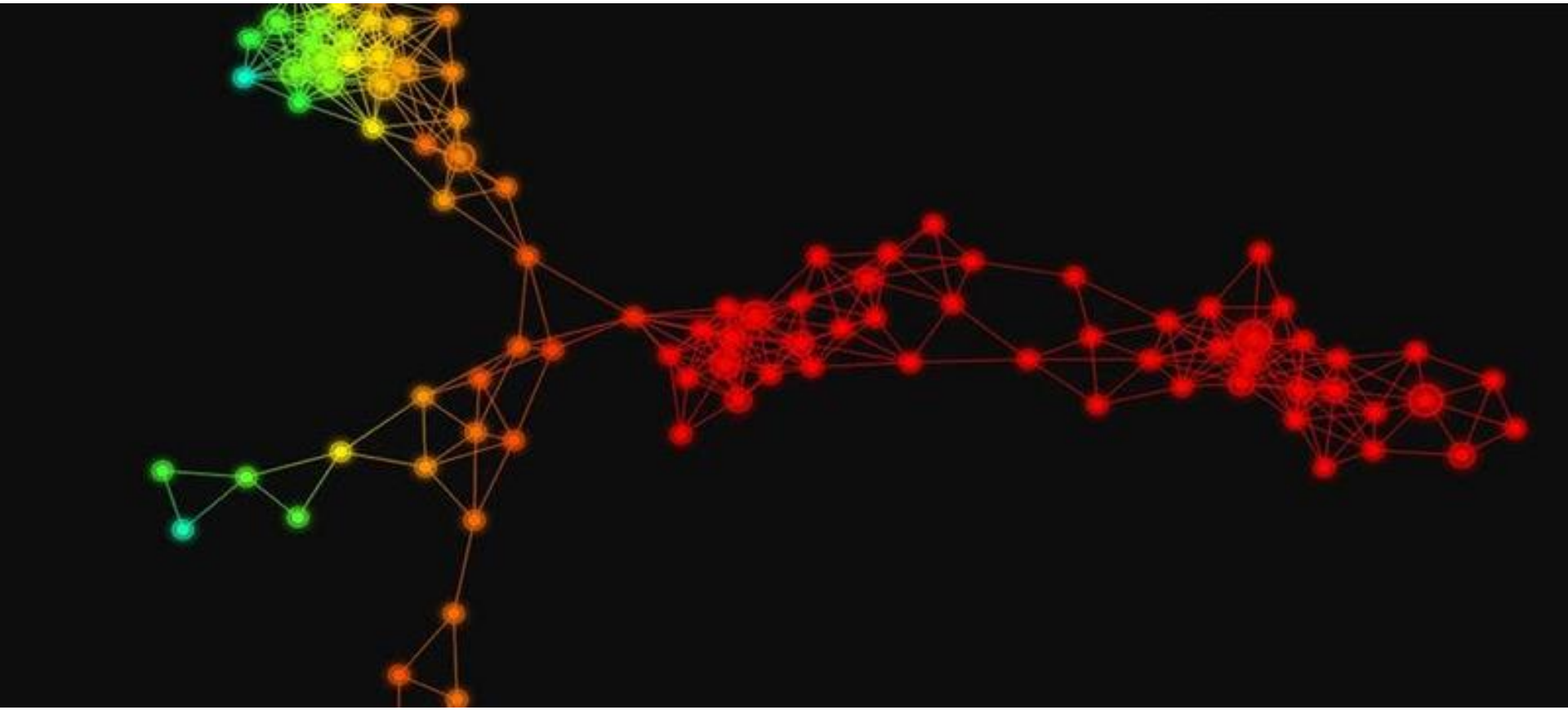
Topological Data Analysis

A Framework for Machine Learning



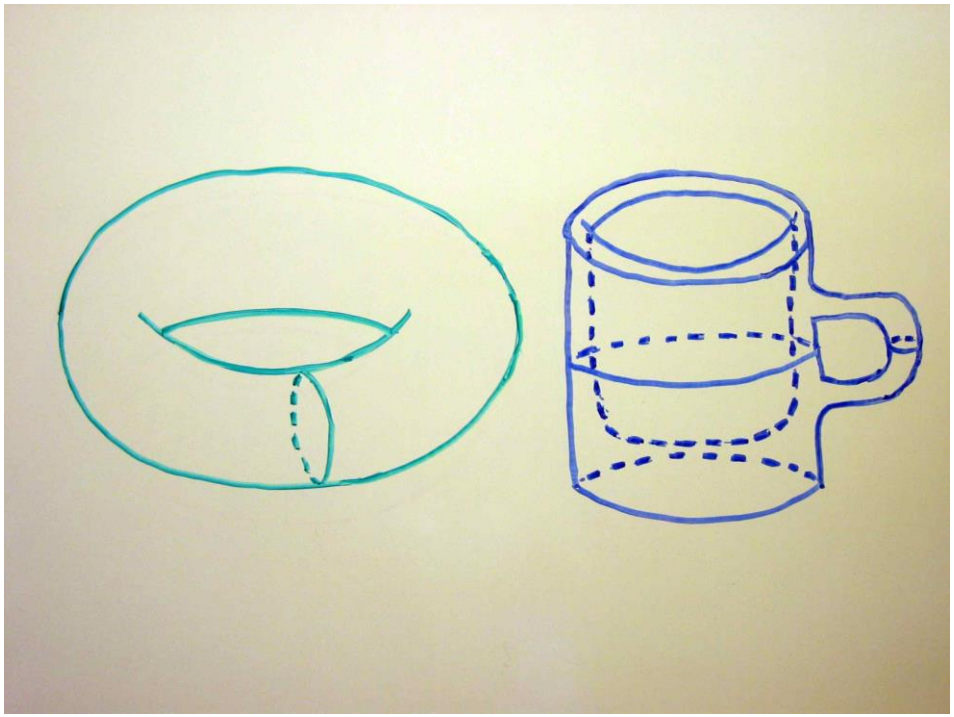
Samarth Bansal (11630) Deepak Choudhary (11234)

Motivation

Named Most Innovative Company in Big Data

*"**Ayasdi** was started in 2008 to bring a groundbreaking new approach to solving the world's most complex problems after a decade of research at Stanford, DARPA and NSF"*

What is Topology?



Topology is a branch of mathematics from the 1700's that studies continuity and connectivity of objects and spaces, utilizing the shape of data to derive meaning in data.

Data has shape.
Shape has meaning.
Meaning derives value.

Goal of TDA : Understand shape without any pre-conceived model

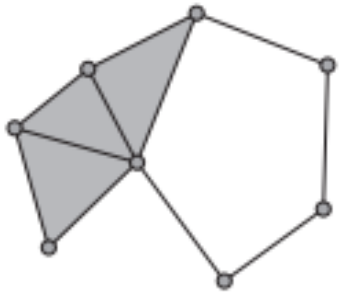
What is TDA?

Extract robust topological features from data and use these summaries for modelling the data.

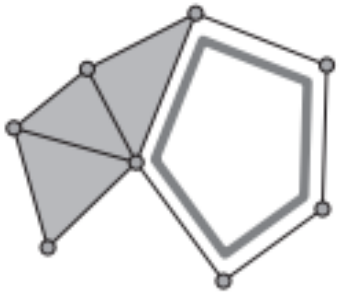
Formal Definition

Given a finite dataset S ⊆ Y of noisy points sampled from an unknown space X, topological data analysis recovers the topology of X, assuming both X and Y are topological spaces.

(a) Points S



(b) Representation K



(c) Invariant

Difference?

Principal Component Analysis (PCA) assumes that X is a linear subspace, a flat hyper plane with no curvature.

ISOMAP assumes that X is intrinsically flat, but is iso-metrically embedded.

Both are instances of **manifold learning**
Assumption : X is a manifold, that is, it is locally Euclidean.

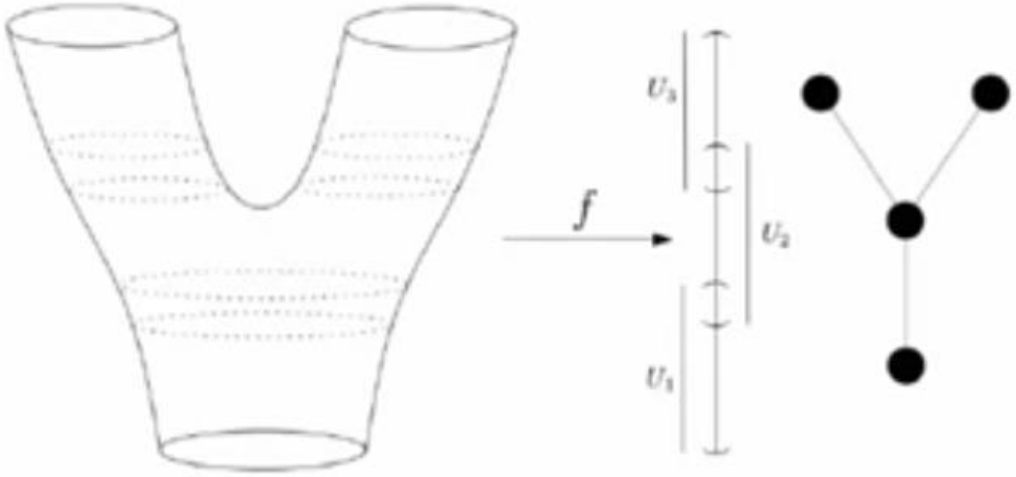
(a) Samples in $\mathbb{R}^2$ (b) Embedding (c) Spiral (d) Geodesic

TDA is **model free** than most statistical methods, since it does not use an a priori linear or algebraic model for the data, rather relies only on measures of similarity.

NO Assumption

Step 1

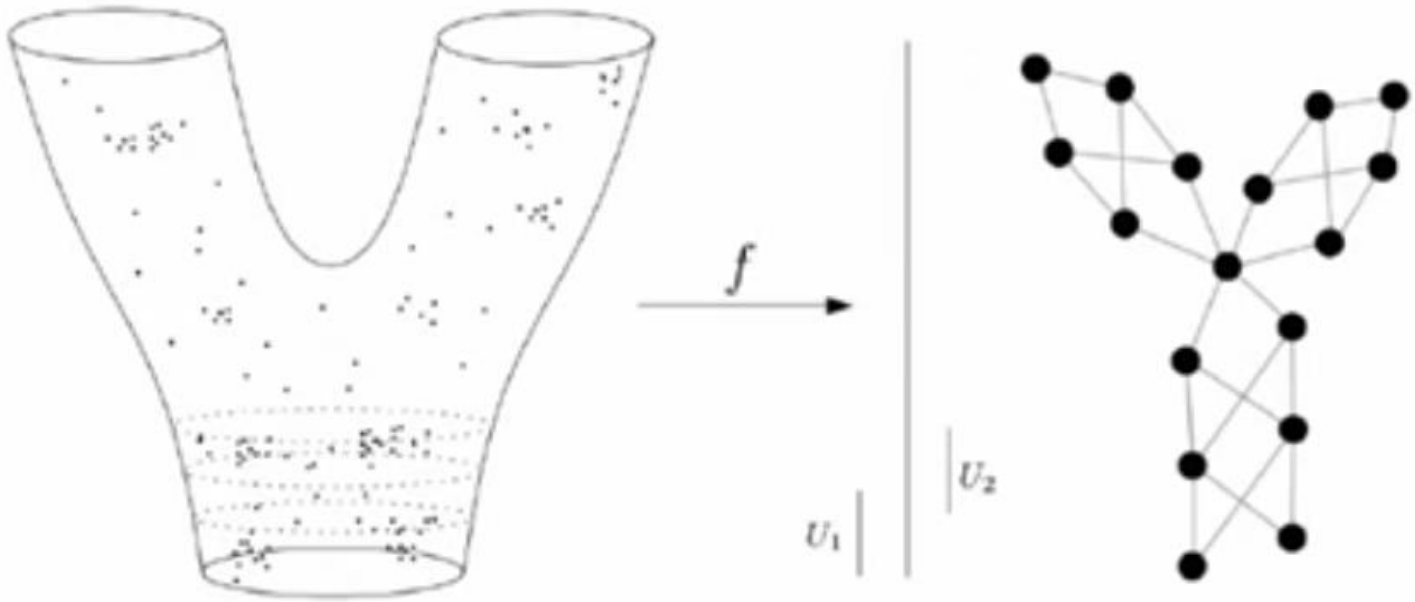
- Replace points in the range with an open covering of the range.
- Connect nodes when their corresponding sets intersect.



⇒ The output is now a graph.

Slide adopted from Anthony Bak's talk on TDA at Stanford Univeristy as part of Colloquium on Computer Systems Seminar Series (EE380)

Step 2: Clustering as π_0



- Nodes are clusters of data points
- Edges represent shared points between the clusters

Slide adopted from Anthony Bak's talk on TDA at Stanford Univeristy as part of Colloquium on Computer Systems Seminar Series (EE380)

Lenses: Where do they come from

- Standard data analysis functions
- Geometry and Topology
- Modern Statistics
- Domain Knowledge / Data Modeling

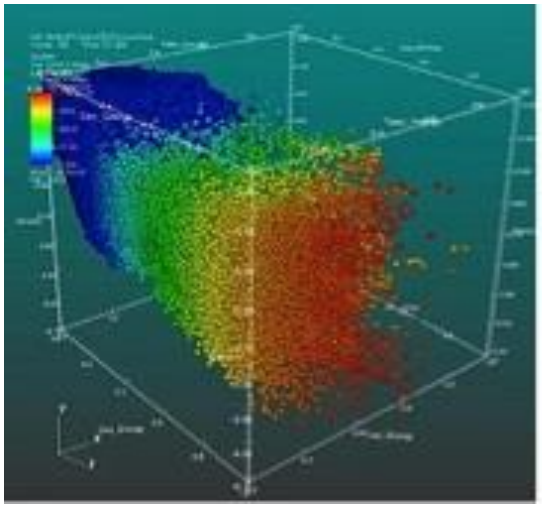
A Non Exhaustive Table of Lenses

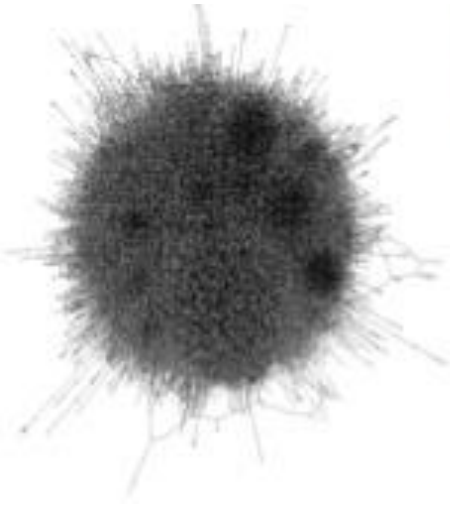
Statistics	Geometry	Machine Learning	Data Driven
Mean/Max/Min	Centrality	PCA/SVD	Age
Variance	Curvature	Autoencoders	Dates
n-Moment	Harmonic Cycles	Isomap/MDS/TSNE	
Density	...	SVM Distance from Hyperplane	
...		Error/Debugging Info	
		...	

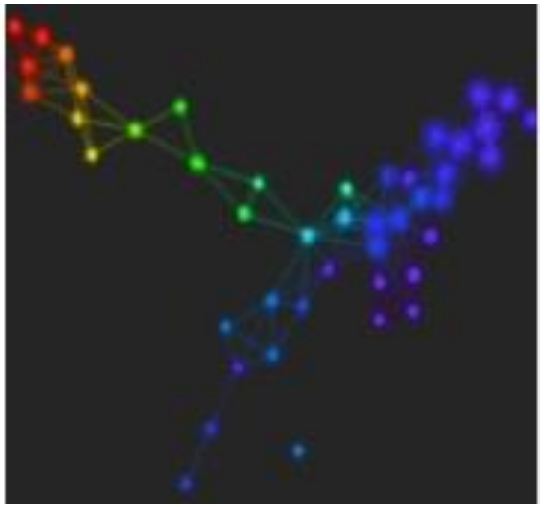
Slide adopted from Anthony Bak's talk on TDA at Stanford Univeristy as part of Colloquium on Computer Systems Seminar Series (EE380)

Visualization Techniques

Scatterplot methods, PCA, MDS





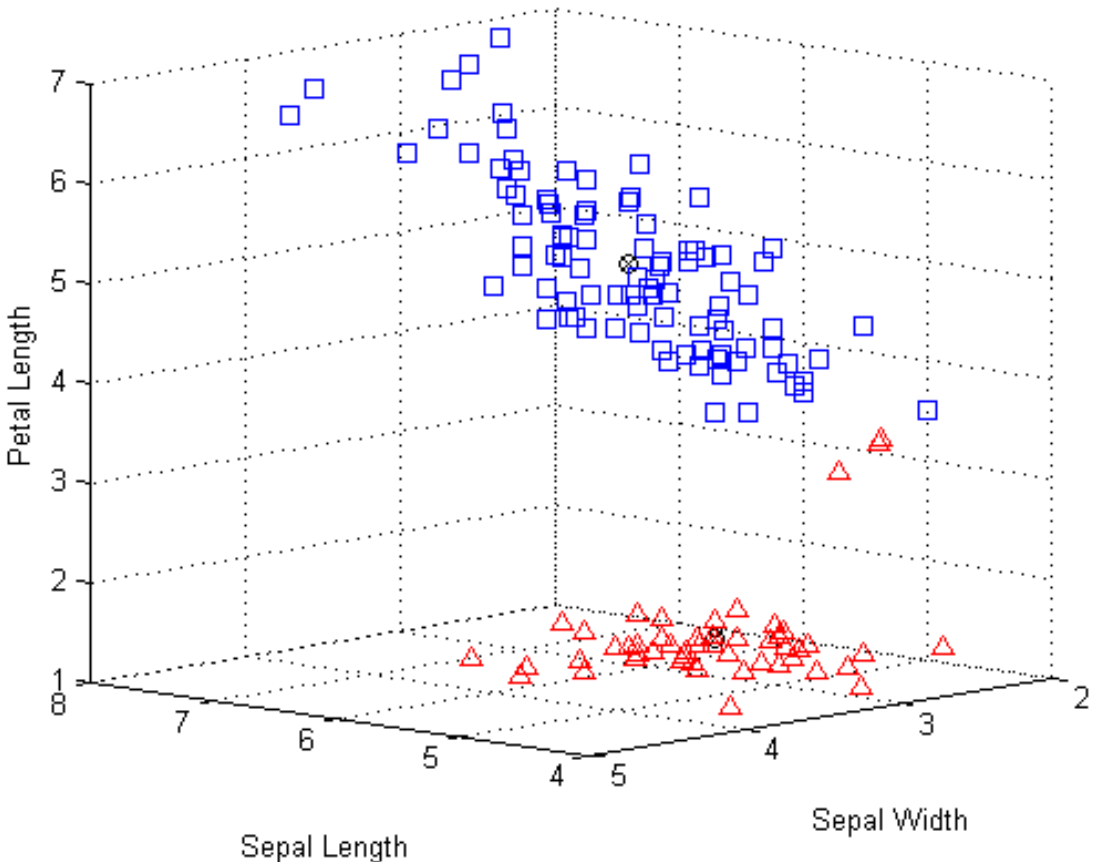


• A topological network represents data by grouping similar data points into nodes, and connecting those nodes by an edge if the corresponding collections have a data point in common.

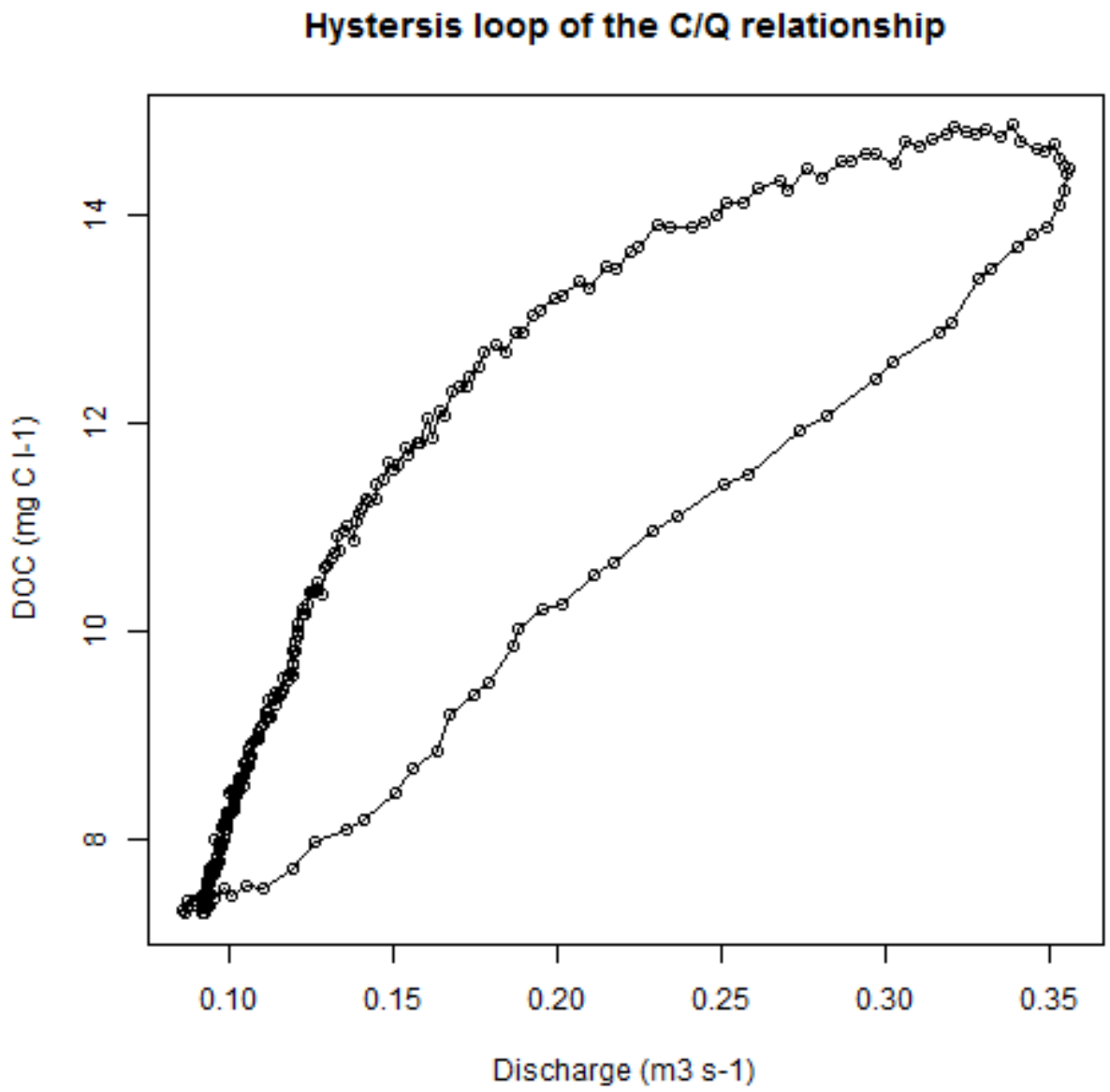
• Because each node represents multiple data points, the network gives a compressed version of extremely high dimensional data.

Cluster Analysis

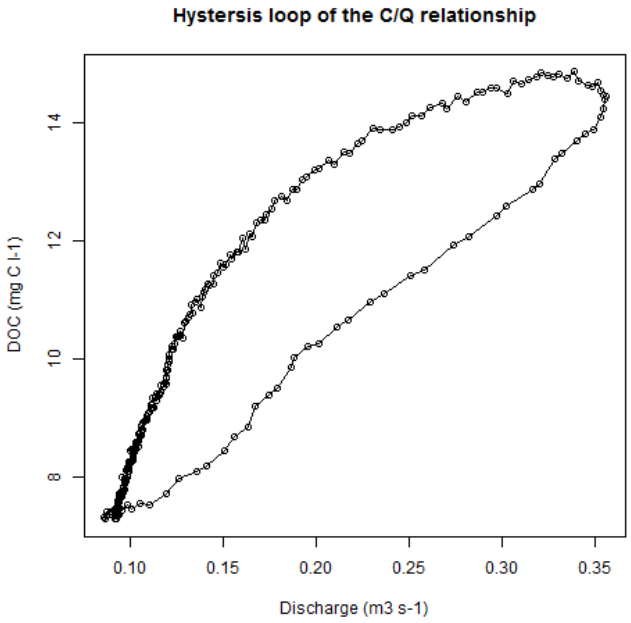
Cluster analysis Goal : Divide a data set up into disjoint groups that have some distinct defining properties, or conceptual coherence.

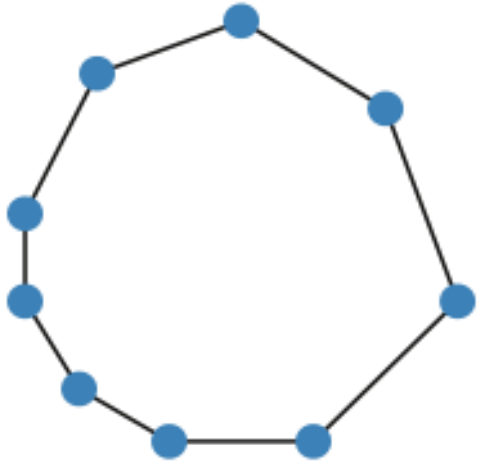


What about this?



TDA!


Hystersis loop of the C/Q relationship
DOC (mg C l-1)
8
10
12
14
0.10
0.15
0.20
0.25
0.30
0.35
Discharge (m3 s-1)



Data transformed into topological networks reveals insights and hidden patterns

The combination of Topological Data Analysis (TDA) with machine-learning automatically creates topological networks revealing statistically significant patterns in complex data

Project Aim

Compare TDA with traditional ML Algorithms

Datasets

Heart Disease Data Set
UCI Machine Learning – 303 Instances, 75 Attributes

Breast Cancer Wisconsin (Original) Data Set
UCI Machine Learning – 699 Instances, 10 Attributes

References

- *Gunnar Carlsson,2009, Bulletin (New Series) of The American Mathematical Society, Volume 46, Number 2, April 2009, Pages 255–308*
- *Lum, P.Y.et al. Extracting insights from the shape of complex data using topology. Sci. Rep. 3, 1236; DOI:10.1038/srep01236 (2013).*
- *Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. Monica Nicolaua, Arnold J. Levineb, and Gunnar Carlssona, Department of Mathematics, Stanford University, Stanford, CA 94305; School of Natural Sciences, Institute for Advanced Study, Princeton, NJ 08540; and Ayasdi, Inc., Palo Alto, CA 94301*