Topology and Data

PSB Workshop
Big Island of Hawaii
Jan.4-8, 2015

Outline

Part I

- Persistent homology & applications

Part II

- TDA & applications

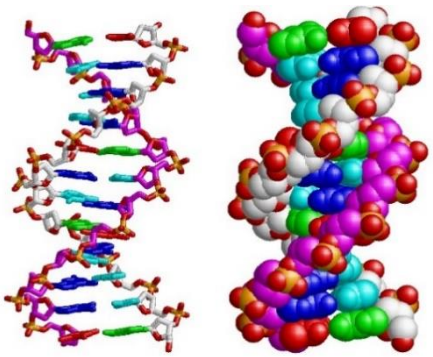
Part I

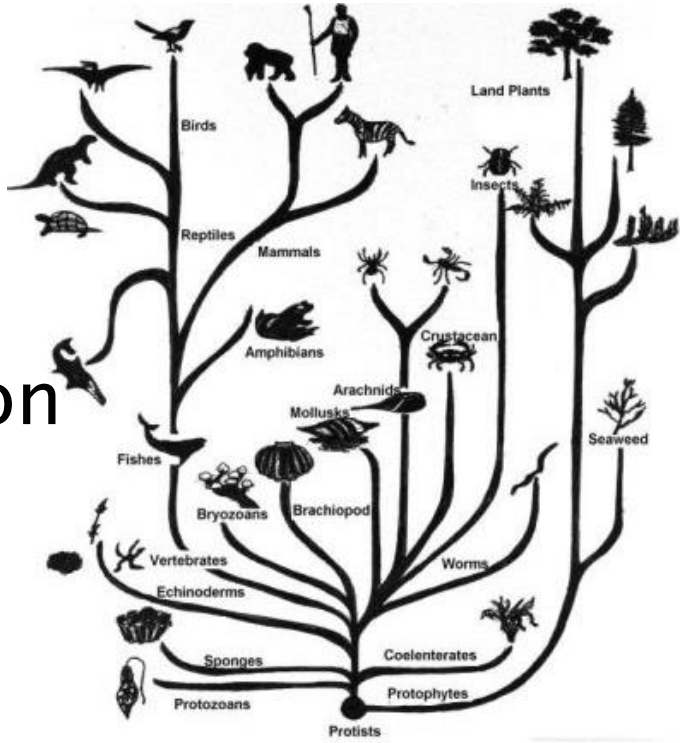
Part I Outline

- Motivating problem from microbiology
- Algebraic Topology
- 2 Examples of applications
  1. Topology of Viral Evolution
  2. Cancer gene expression
- Available software
- Questions

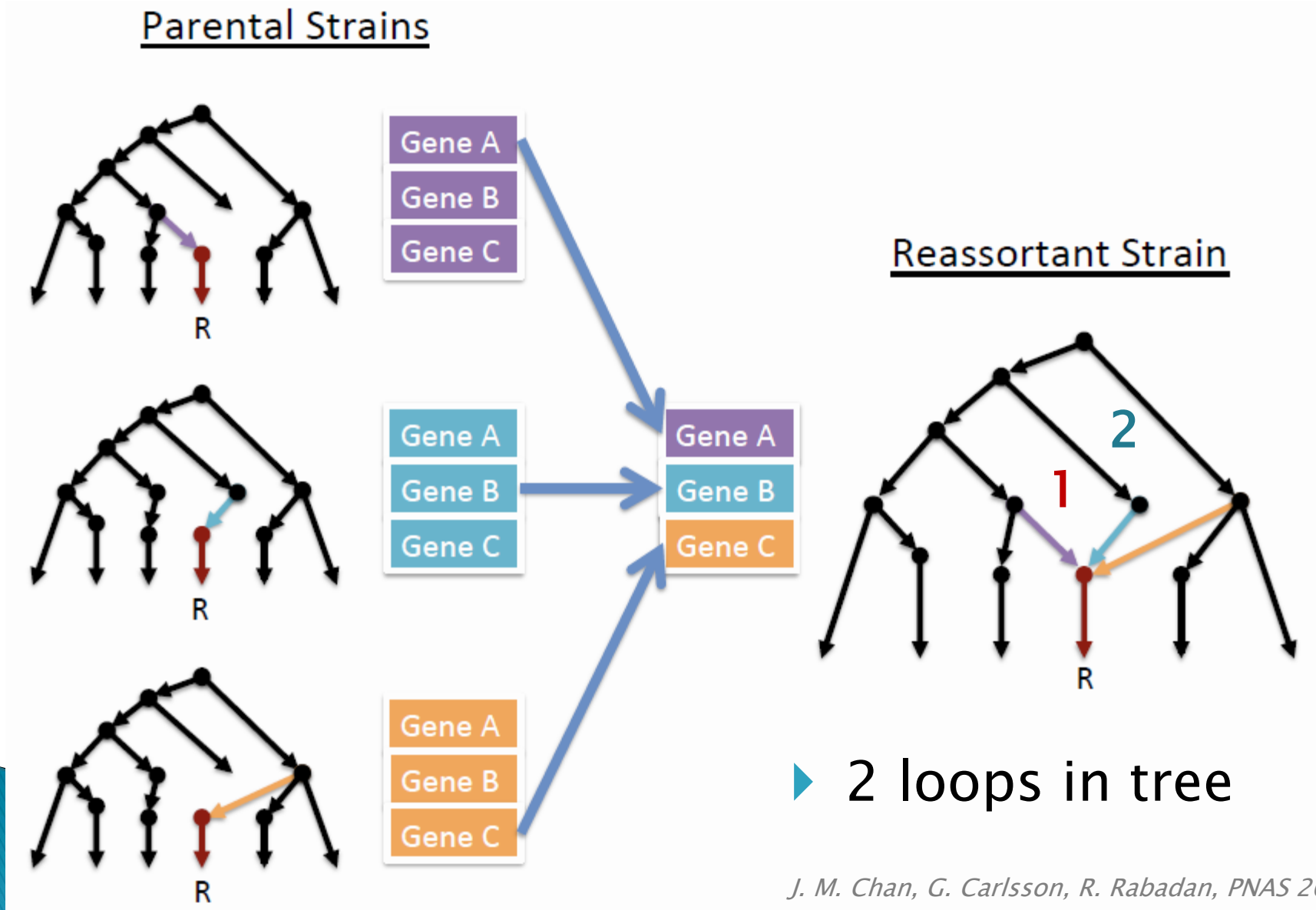
Genes & Evolution

- DNA provides genetic material for all living organisms on Earth we know
- Tree of Life – reconstruction of *vertical* evolution
- No gene exchange among branches
- However, horizontal evolution is also present especially in microorganisms





Genes & Evolution



- 2 loops in tree

J. M. Chan, G. Carlsson, R. Rabadan, PNAS 2013

Genes & Evolution

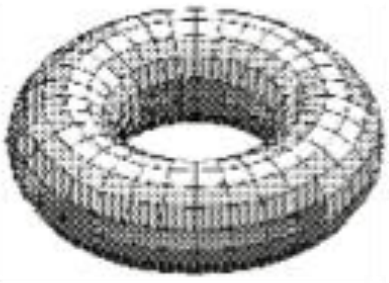
- The tree structure can capture vertical evolution only
- What structure can capture both horizontal and vertical evolutions?

- 2013 PNAS – “Topology of Viral Evolution”, Joseph Minhow Chan, Gunnar Carlsson, and Raul Rabadan

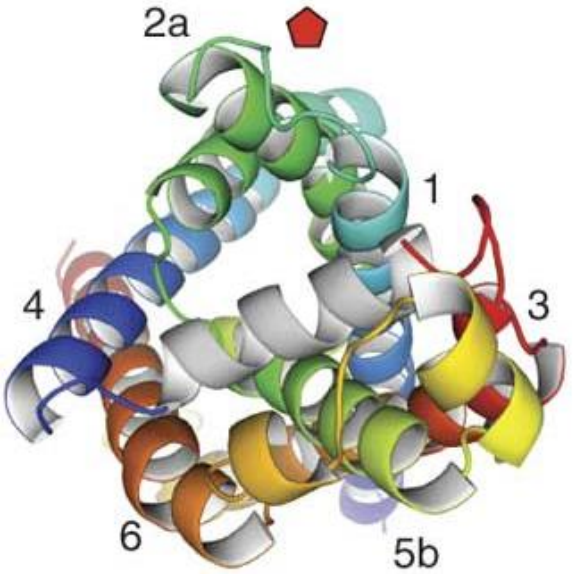
Topology (math)

Important note: do not confuse biological and mathematical terms of *"topology"* and *"homology"*

Topology (math)



Topology (molecular)



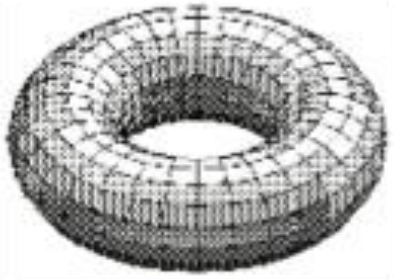
Topology: Brief

- Topology studies notions of *shape* and as long as pairwise distance are well defined, one can study topology of data
- Clustering, graphs, networks inform about *local* properties of data (connectivity, density, etc.)
- Topology, and its subset persistent homology, informs us about *global* properties of data (# of connected components, enclosed voids, etc.)

$\beta_0 = 1$

$\beta_1 = 0$

$\beta_2 = 1$



$\beta_0 = 1$

$\beta_1 = 2$

$\beta_2 = 1$

Persistent Homology: The Method

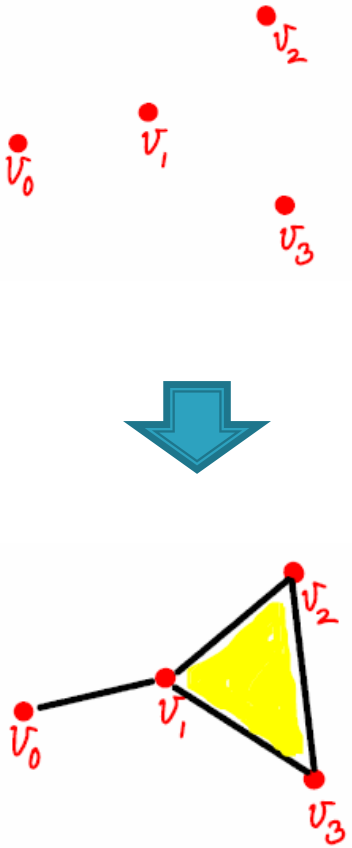
Barcoding: find topological invariants in cloud data

1. Data transformed in high-dimensional space using pairwise distance
2. Construct family of nested ***simplicial complexes***, indexed by a proximity parameter
3. Encode the persistent homology of a data set in the form of a parameterized version of a ***Betti number***: a ***barcode***
4. Analyze the barcodes

Simplicial Complexes

Crash course in Computational Algebraic Topology

- Use data points as vertices of a graph whose edges are determined by proximity (vertices within some specified distance)
- Graph serves as a scaffold for a *simplicial complex* – a structure of points, line segments, triangles, tetrahedra, and etc.

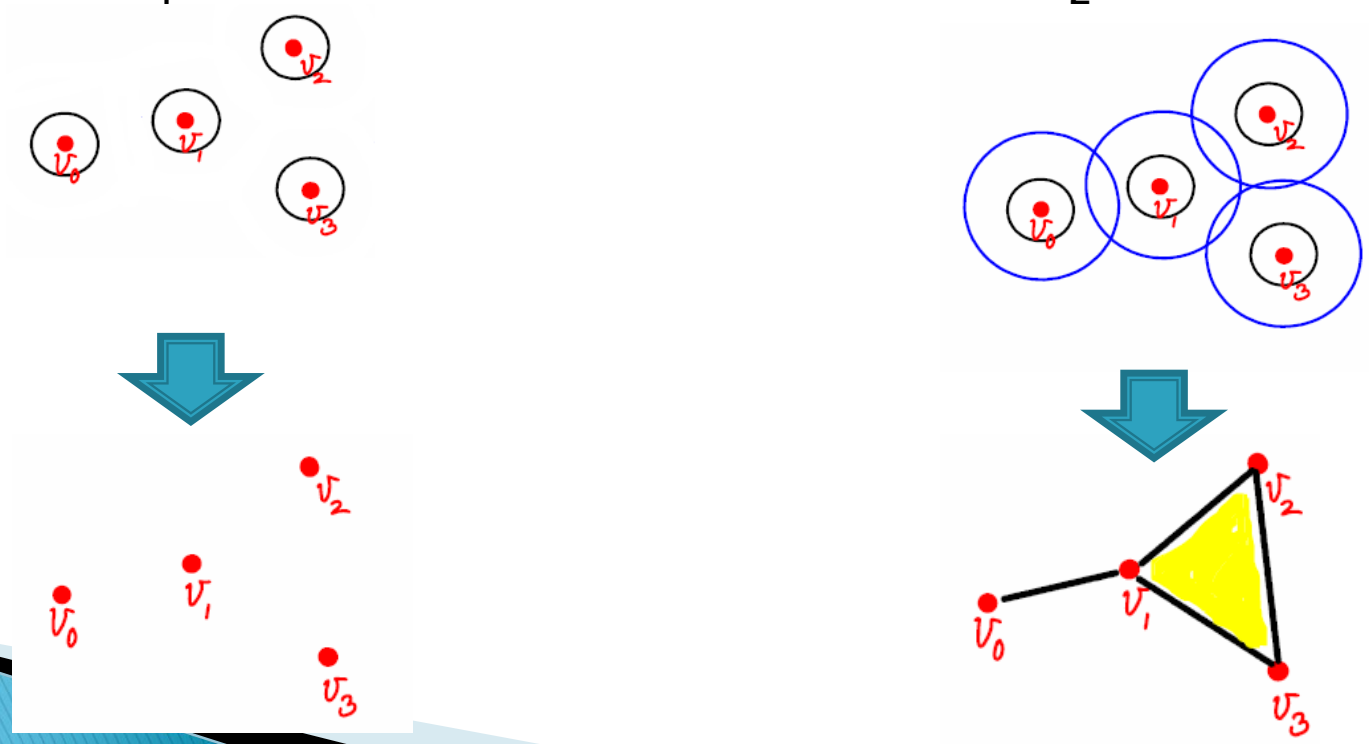


Simplicial Complexes

Crash course in Computational Algebraic Topology

One of the most commonly used simplicial complexes is Vietoris-Rips complex

r_1 < r_2

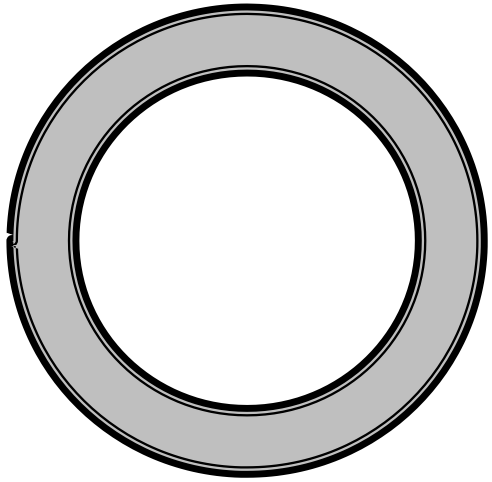


Persistent Homology

Crash course in Computational Algebraic Topology

- What is an ideal *r*? Which holes are “real” and which are “noise”?
- Persistence is a rigorous response to this problem
- Topological features which ***persist*** over a significant parameter range are to be considered as signal
- Short-lived features are to be considered as noise
- Topological features are encoded as ***barcodes*** and ***Betti numbers***

Persistent Homology: Betti #'s



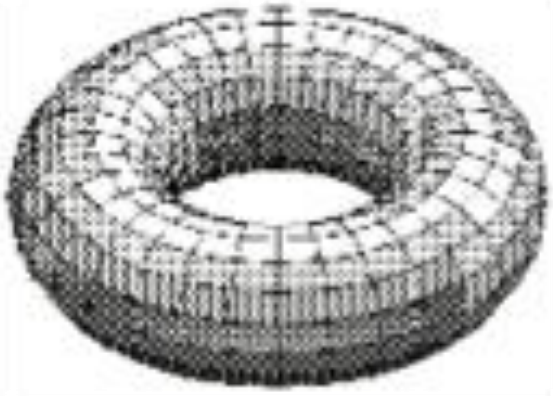
$\beta_0 = 1$

$\beta_1 = 1$

$\beta_2 = 0$

Informally:

- β_0 – # of connected components
- β_1 – # of loops
- β_2 – # of voids



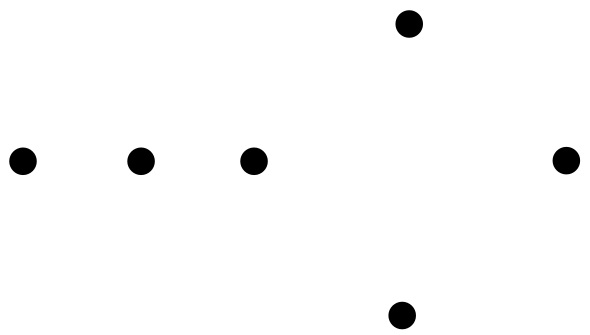
$\beta_0 = 1$

$\beta_1 = 2$

$\beta_2 = 1$

Persistent Homology: Barcodes

- Example: 6 points in $\mathbb{R}^2$, filtration function – Euclidean distance

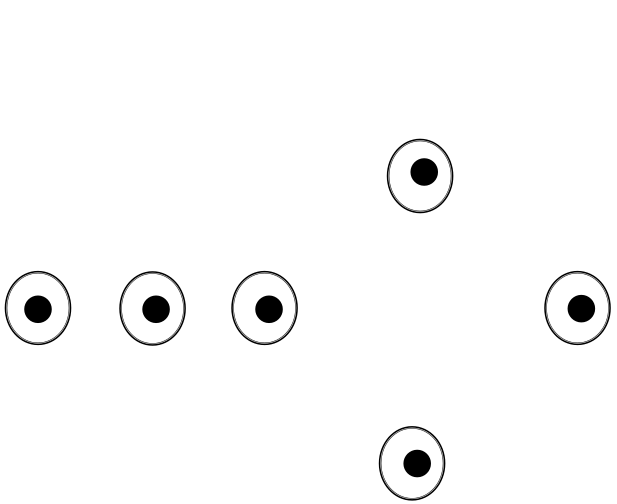


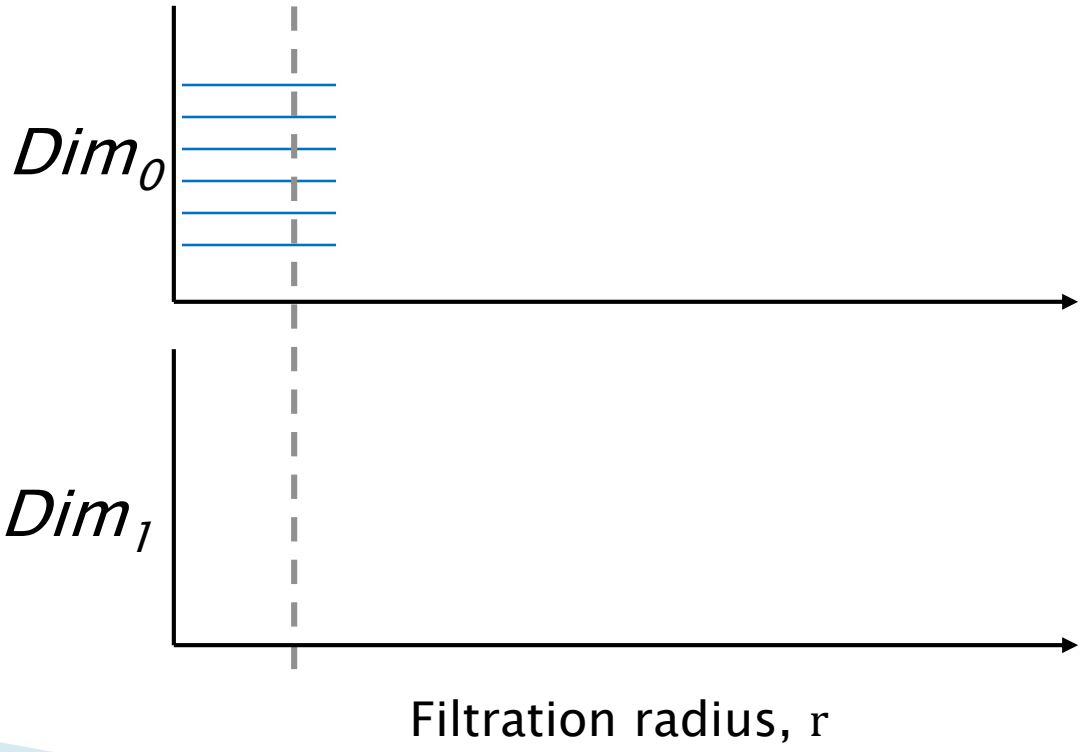
Persistent Homology: Barcodes

- Each of 6 data points corresponds to a barcode
  - *r* is small

$\beta_0 = 6$

$\beta_1 = 0$



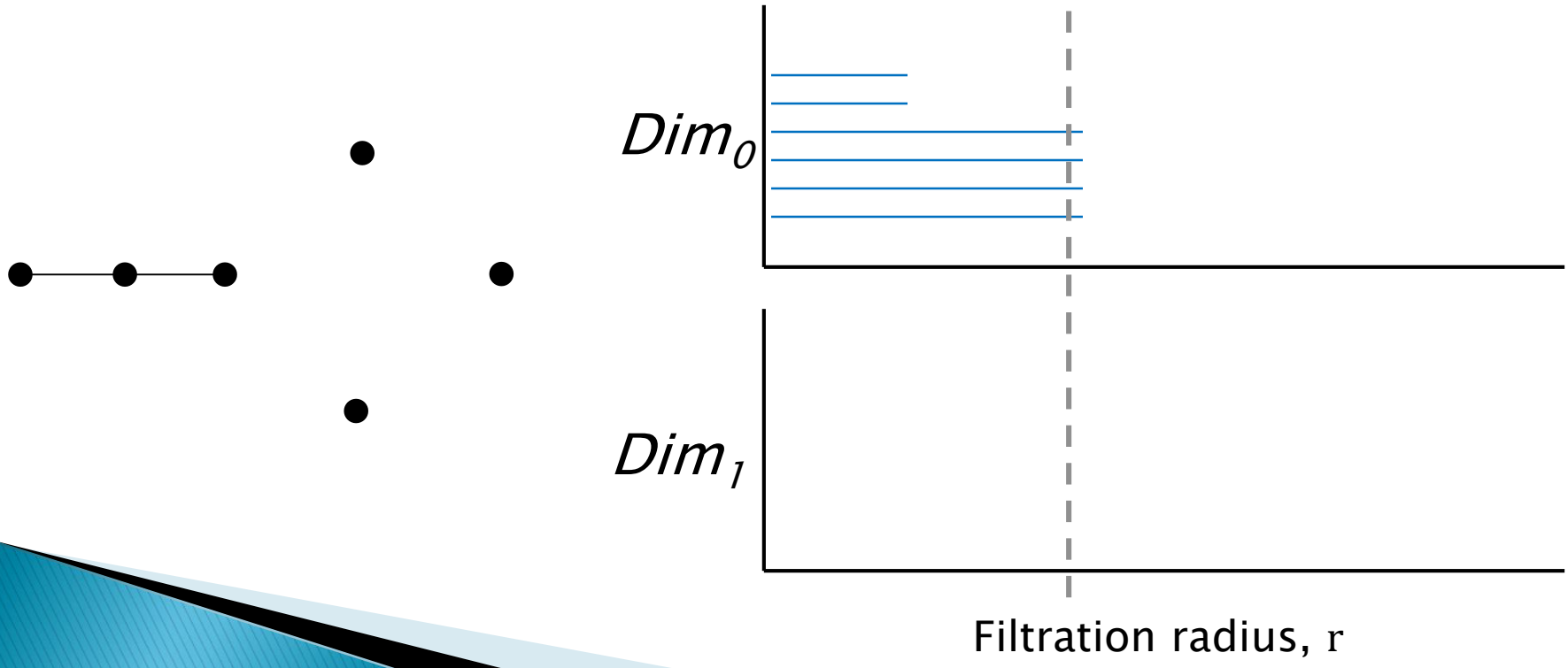


Persistent Homology: Barcodes

- Filtration radius increases “killing” 2 components

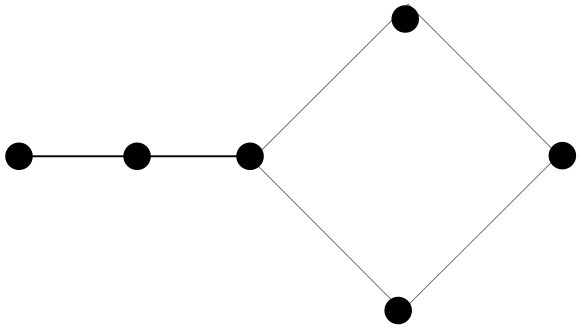
$\beta_0 = 6$ $\beta_0 = 4$

$\beta_1 = 0$ $\beta_1 = 0$



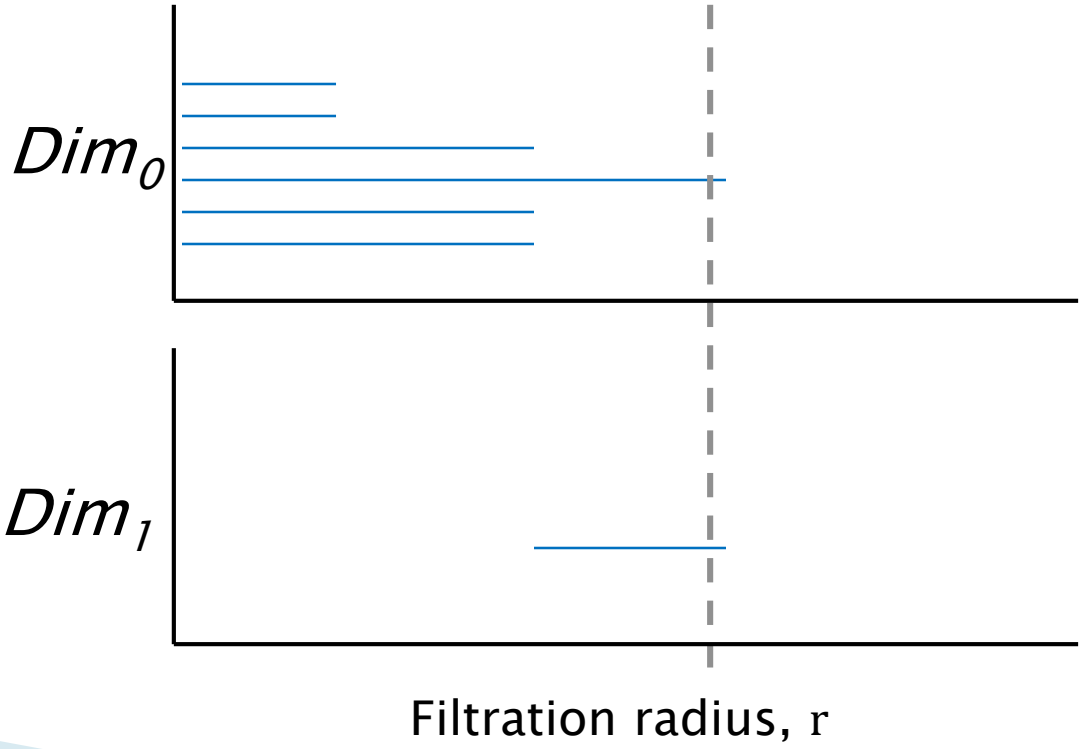
Persistent Homology: Barcodes

- Increasing radius further creates a loop in Dim_1



$\beta_0 = 6$ $\beta_0 = 4$ $\beta_0 = 1$

$\beta_1 = 0$ $\beta_1 = 0$ $\beta_1 = 1$

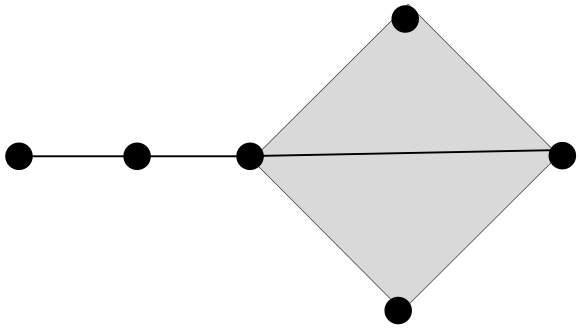


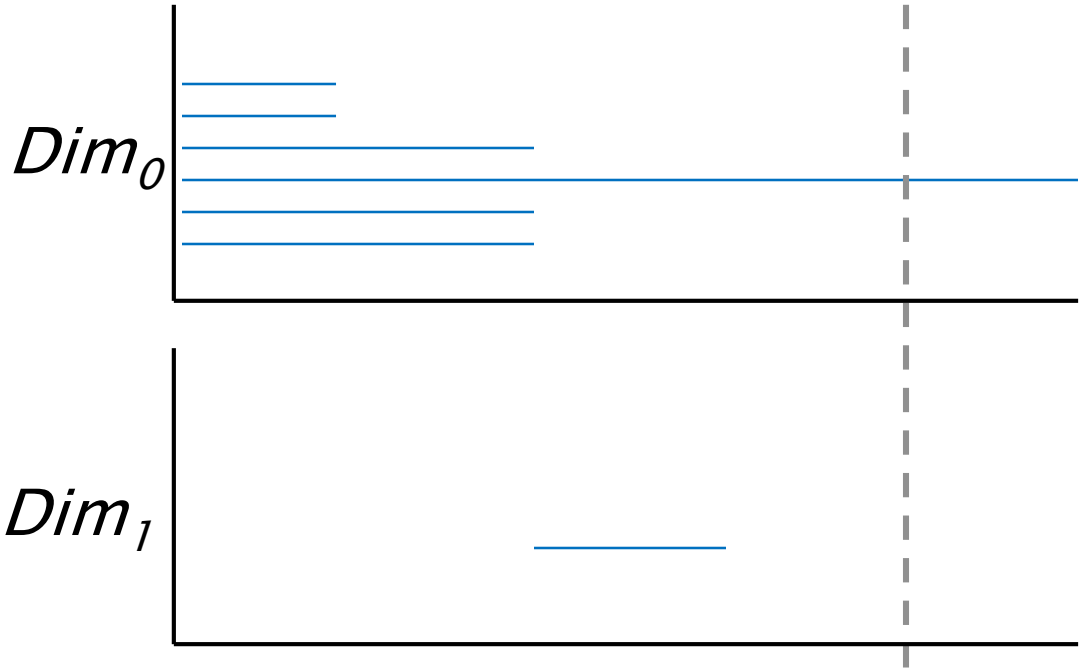
Persistent Homology: Barcodes

- The loop close, 1 connected component persist

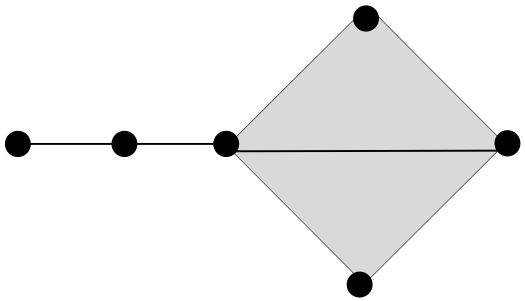
$\beta_0 = 6$ $\beta_0 = 4$ $\beta_0 = 1$ $\beta_0 = 1$

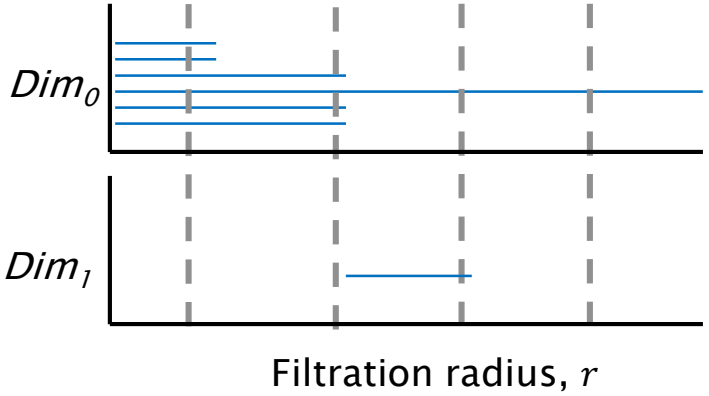
$\beta_1 = 0$ $\beta_1 = 0$ $\beta_1 = 1$ $\beta_1 = 0$





Persistent Homology: Barcodes



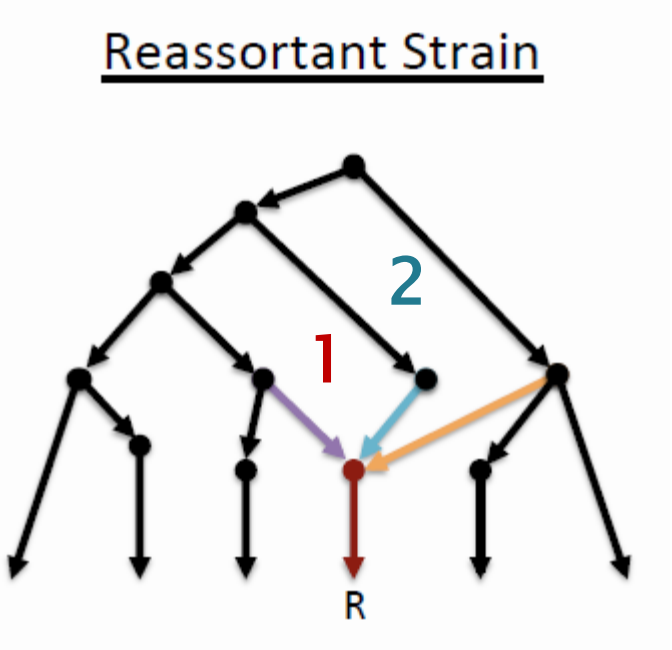


- Evolution of barcodes and Betti numbers provide bases for analysis of data

Questions?

Topology of Viral Evolution

- What structure can capture both horizontal and vertical evolutions?
- We would like to capture the loops as well as vertical evolution

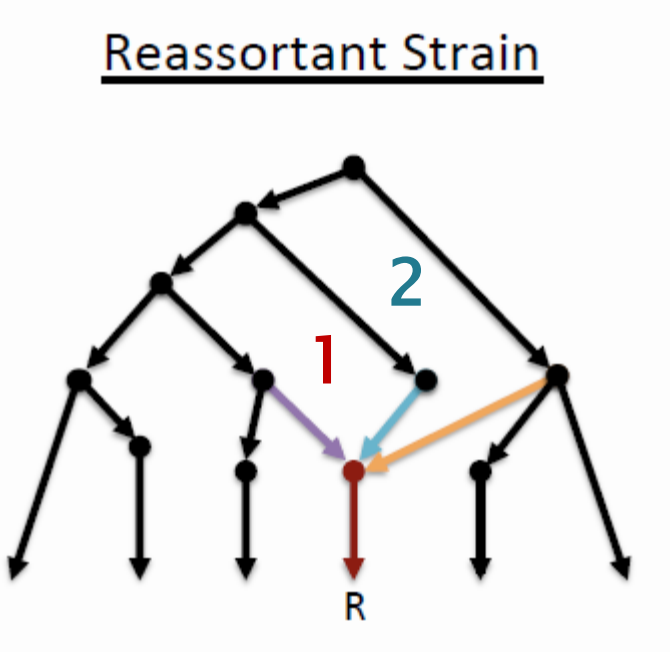


- 2 loops in tree

J. M. Chan, G. Carlsson, R. Rabadan, PNAS 2013

Topology of Viral Evolution

- ***Persistent homology*** mathematically formalizes that
- The barcodes in Dim.0 (β_0) provide information about vertical evolution
- The barcodes in Dim.1 (β_1) inform about horizontal evolution
- Persistent homology was applied to RNA viruses

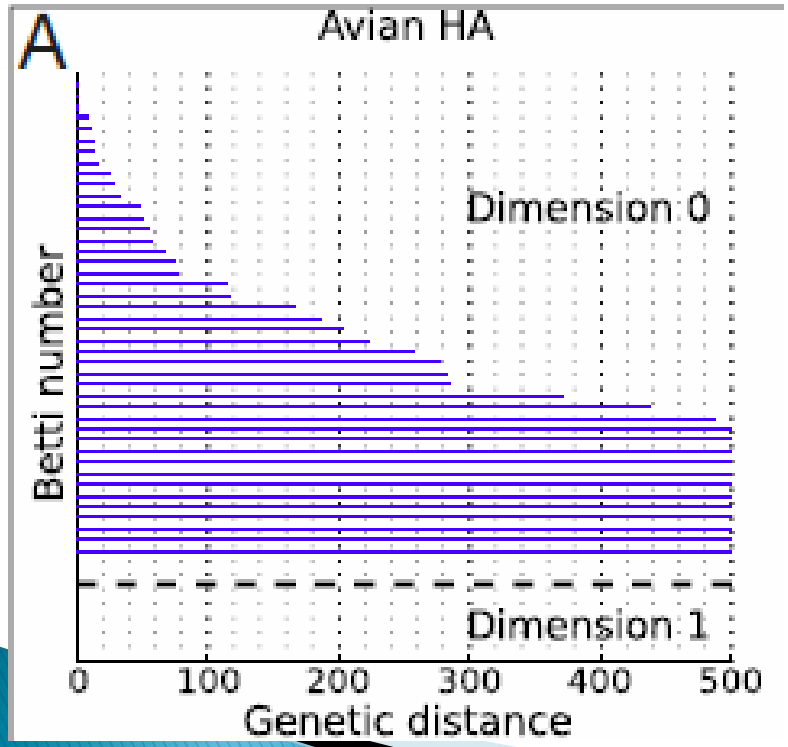


J. M. Chan, G. Carlsson, R. Rabadan, PNAS 2013

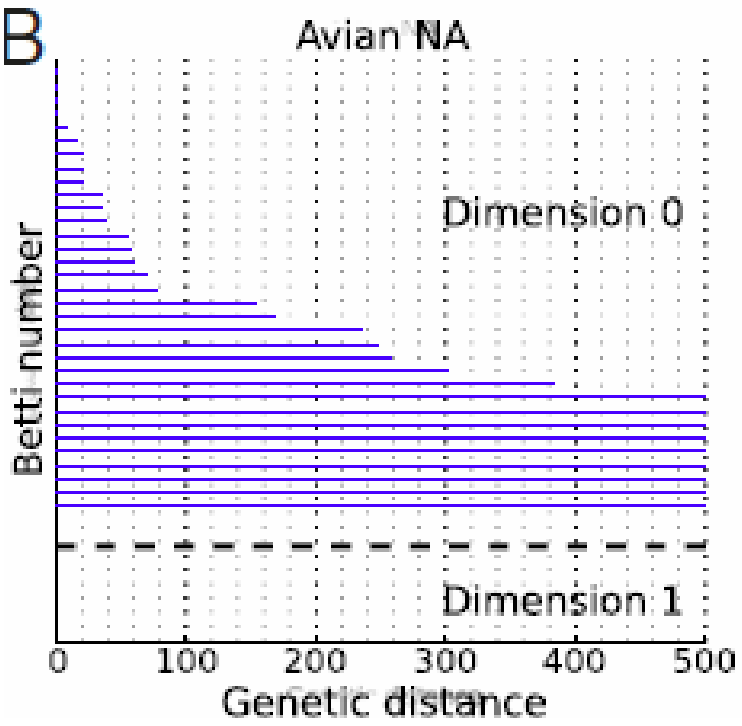
Avian Influenza

- Individual genes would not produce high dimensional topology ($\beta_n = 0, for\ n \geq 1$)

HA – hemagglutinin



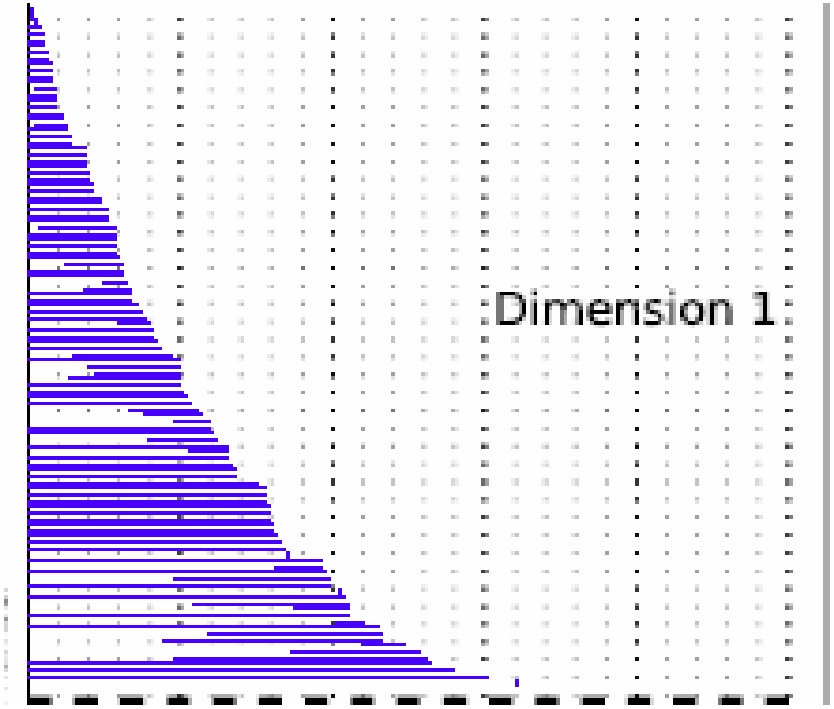
NA – neuraminidase

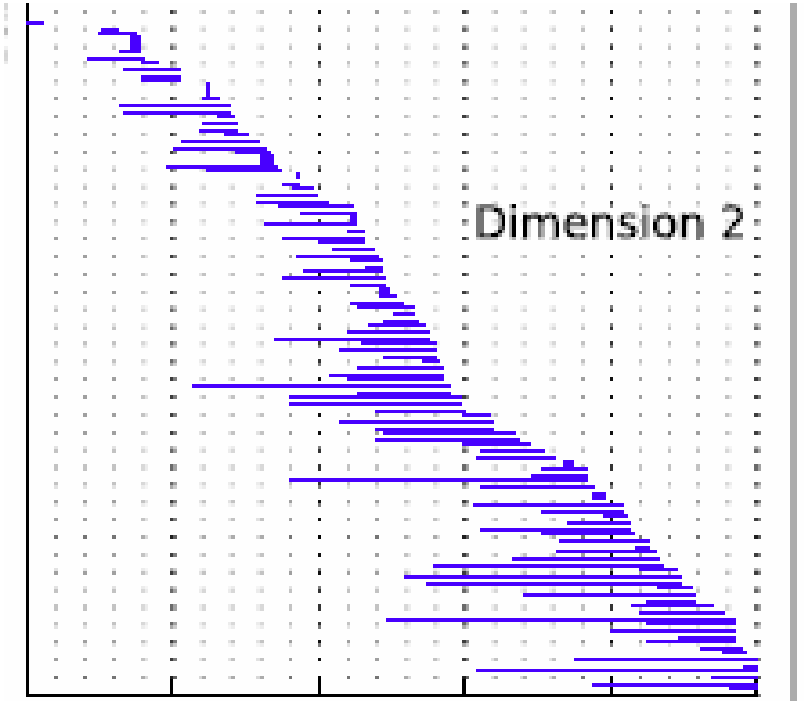


J. M. Chan, G. Carlsson, R. Rabadan, PNAS 2013

Avian Influenza

- However, concatenation of HA and NA produces:



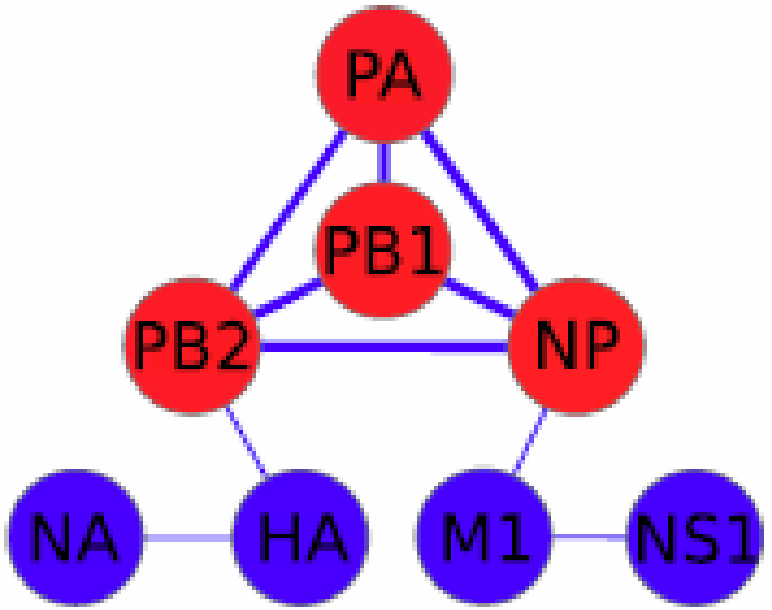


- Also used other virus segments–
  PA, PB1, PB2, NP, M1, NS1

J. M. Chan, G. Carlsson, R. Rabadan, PNAS 2013

Avian Influenza

- Use barcodes of Dim. 1 to determine pattern of gene segment association
- Statistically significant configuration of four cosegregating segments – PB2, PB1, PA, and NP

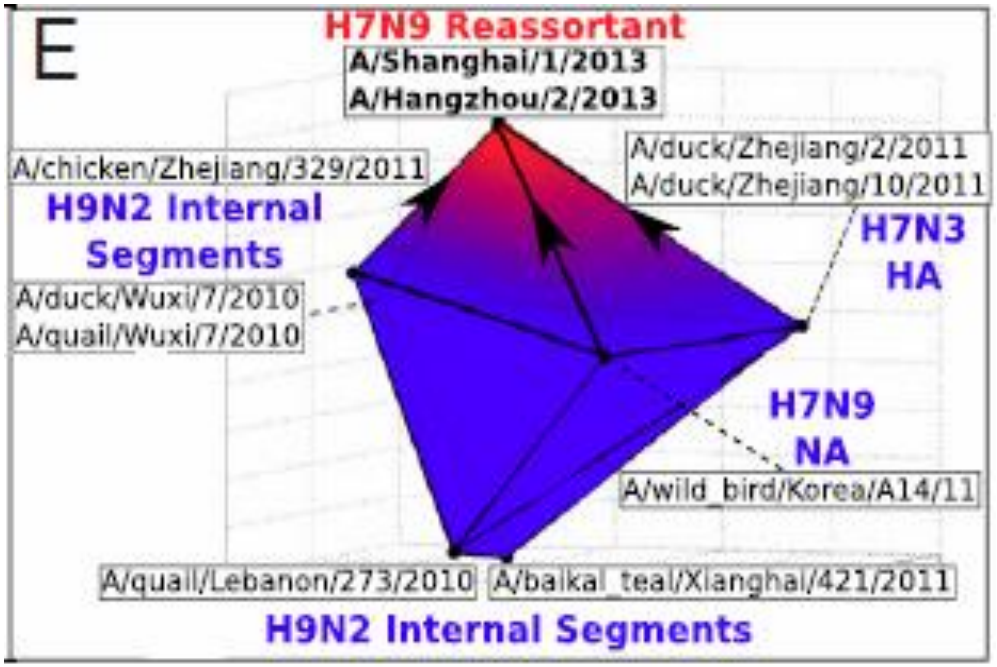


  - Compute probability p_{ij} that two segments co-segregate given that we observe $\#\beta_{1ij}$ events in a total of $\#\beta_1$

J. M. Chan, G. Carlsson, R. Rabadan, PNAS 2013

Avian Influenza

- Using similar methodology but on β_2 (Dim.2)
  - Get 3D polytope



  - H7N9 Avian influenza triple reassortment
    - Supported by previous studies

J. M. Chan, G. Carlsson, R. Rabadan, PNAS 2013

Summary

- Persistent homology allowed:
  1. Fast extraction of large-scale patterns from genomic data
  2. Capture succinctly the history of complex genetic exchanges
  3. Reconstruct both vertical and horizontal evolutionary events at the same time

Questions?

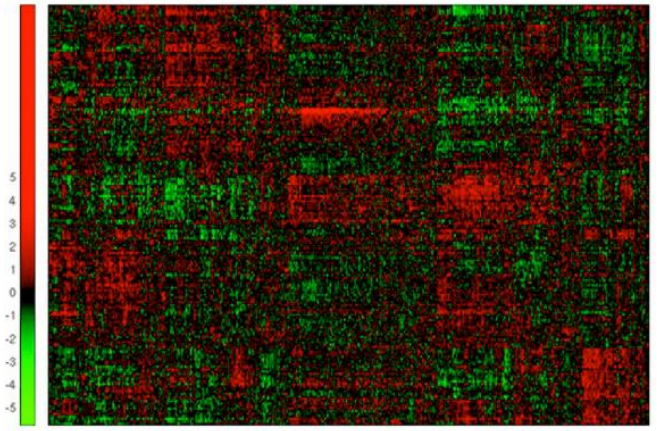
Topology of Cancer (2015)

PSB 2015 – "Topological Features in Cancer Gene Expression Data"

Svetlana Lockwood, Bala Krishnamoorthy

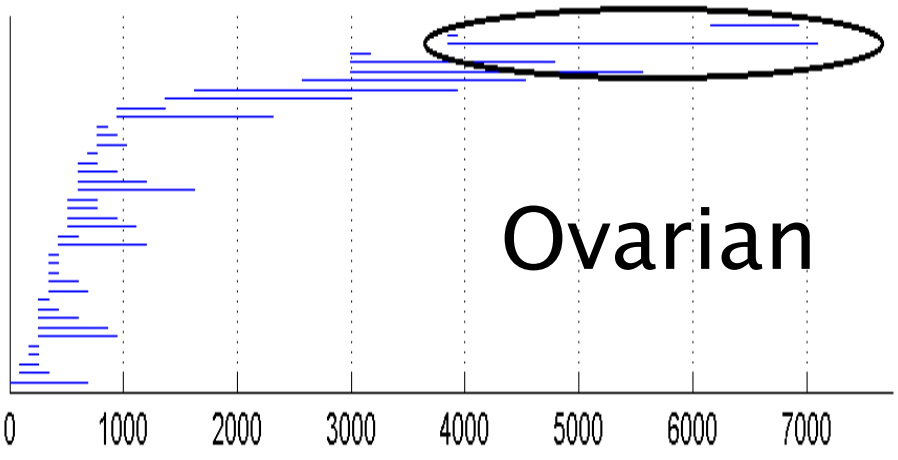
Topology of Cancer (2015)

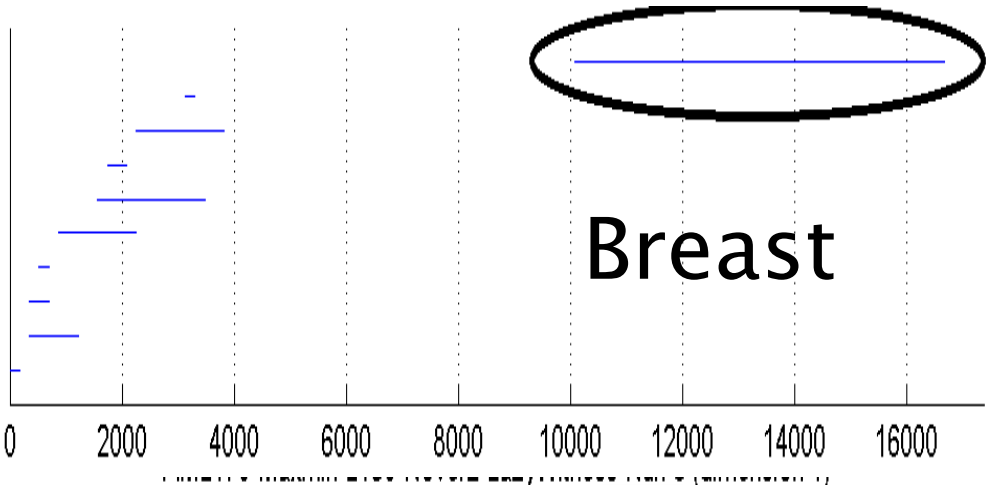
- Microarrays provide expression of tens of thousands of genes
- Allows to search for cancer biomarkers
- Challenge: select a set of genes relevant to cancer

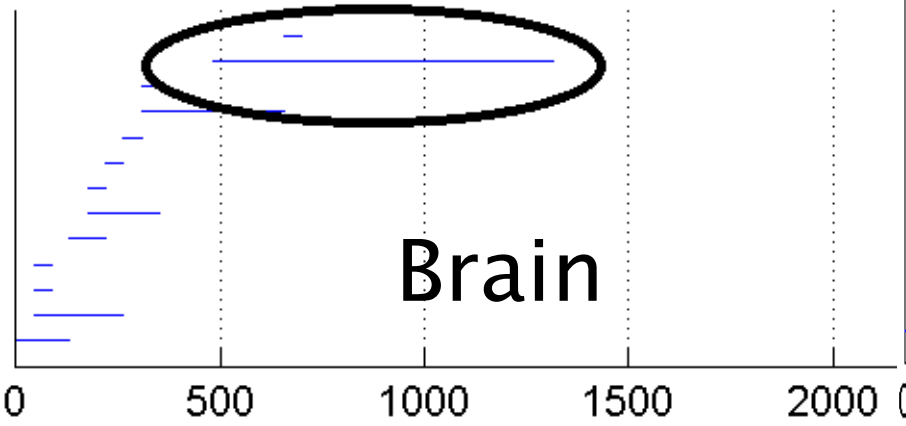


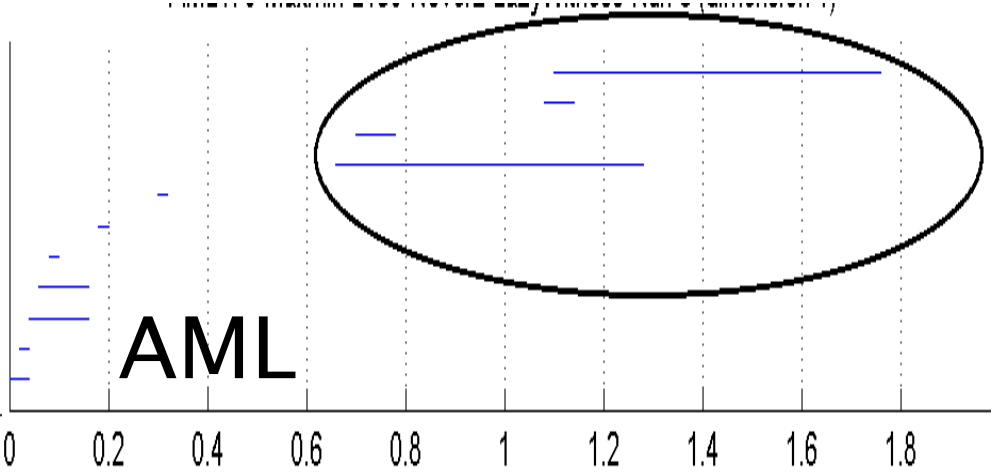
- Hypothesis: ***geometric*** connectedness of genes in loops may imply ***functional*** connectedness

Topology of Cancer (2015)









Topology of Cancer (2015)

- Analyzed 5 cancer datasets

Dataset	Total Genes	#Loops	#Genes in Loops
Brain	46201	1	13 (9)
Breast	54613	1	10 (8)
Ovarian	54613	1	17 (9)
AML188	54613	2	33 (14)
AML170	12558	2	19 (10)

Topology of Cancer (2015)

- Many of genes in loops implicated in cancer

Gene	Dataset	Description
CAV1	Brain	tumor suppressor gene
RPL36	Brain	prognostic marker in hepatocellular carcinoma
RPS11	Breast	downregulation in breast carcinoma cells
FTL	Breast	prognostic biomarkers in breast cancer
LDHA	Ovarian	overexpressed in tumors, important for cell growth
GNAS	Ovarian	biomarker for survival in ovarian cancer
LAMP1	AML170	regulation of melanoma metastasis
PABPC1	AML170	correlation with tumor progression
HLF	AML188	promotes resistance to cell death
DTNA	AML188	induces apoptosis in leukemia cells

Questions?

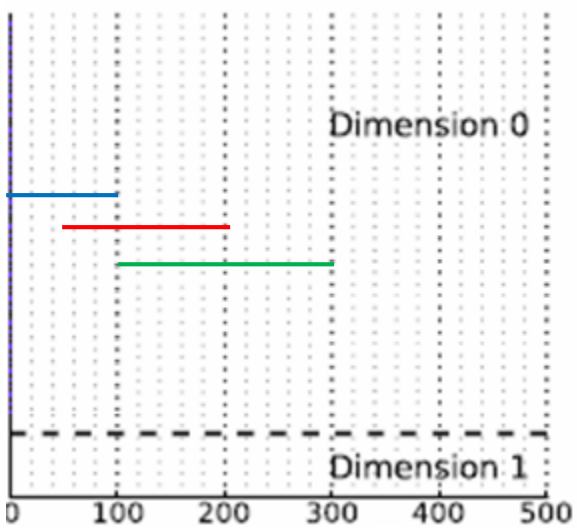
Available software

arxiv, 2015 – “A Roadmap For The Computation Of Persistent Homology”

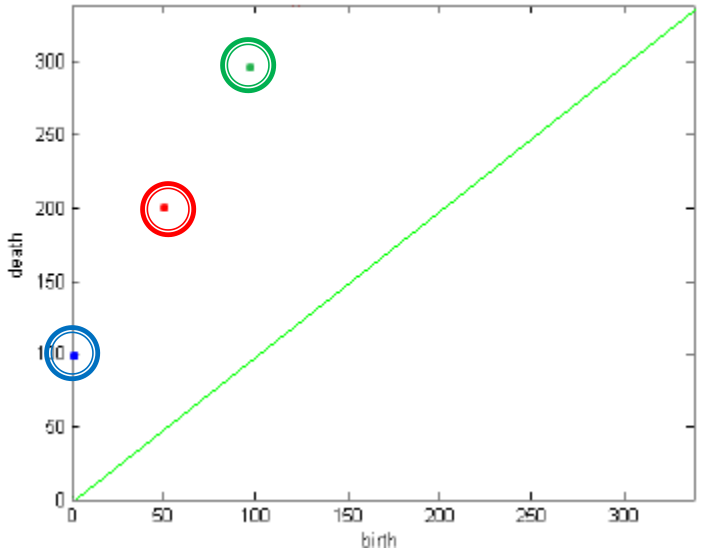
Nina Otter, Mason A. Porter, Ulrike Tillmann, Peter Grindrod, Heather A. Harrington (Oxford)

Available software

- Two major visualizations – barcodes and dots on birth-death plot



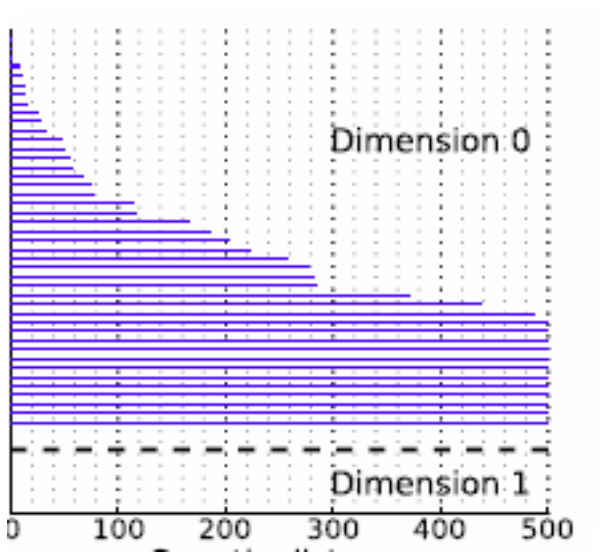
Barcodes by javaPlex



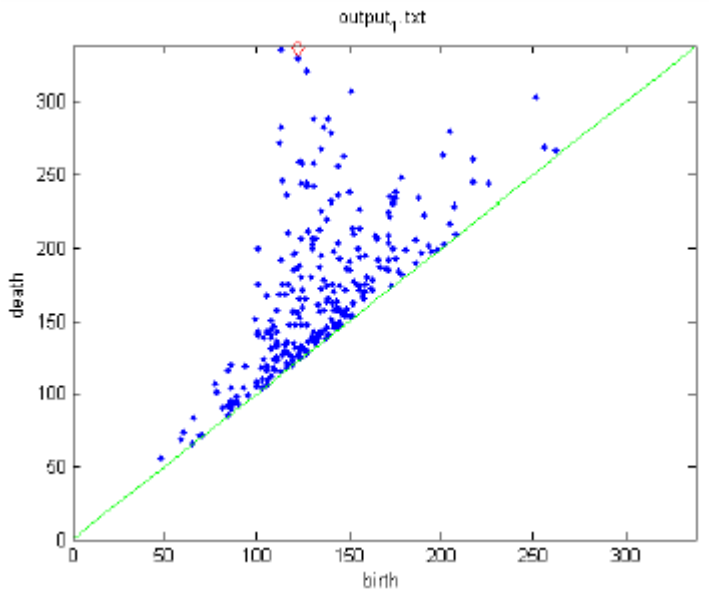
Dots by Perseus

Available software

- This is how they usually look



Barcodes by javaPlex



Dots by Perseus

Available software

A number of open source software is available for computing persistent homology

Software	Installation	Complex	Boundary matrix	Barcodes	Visualization
JavaPlex	✓	✓	✓	✓	✓
Perseus	✓	✓	✓	✓	✓
Dionysus	--	✓	✓	✓	--
DIPHA	--	✓	✓	✓	✓
GUDHI	--	✓	✓	✓	--

arxiv 2015, N. Otter, M. A. Porter, U. Tillmann, P. Grindrod, H. A. Harrington

Available software

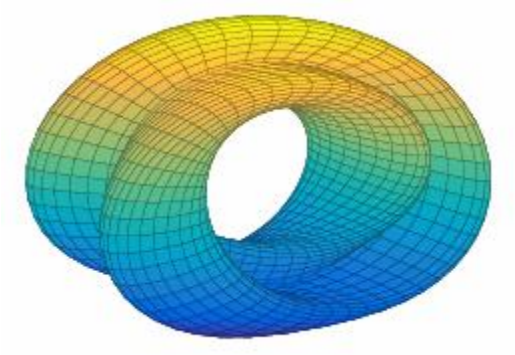


Figure-8 immersion of the Klein bottle

Stanford Dragon

Dataset	C. elegans	Klein	HIV	Dragon 1	Dragon 2
size of complex	4.4×10^6	1.1×10^7	2.1×10^8	1.7×10^8	1.3×10^9
JavaPlex	284	1031	--	--	--
Perseus	542	1974	--	--	--
Dionysus*	513	145	--	4362	--
DIPHA*	39	6	1276	1176	37572
GUDHI	4	11	248	283	3151

* Dual implementation

- CPU time in seconds

arxiv 2015, N. Otter, M. A. Porter, U. Tillmann, P. Grindrod, H. A. Harrington

Available software

Summary:

- **javaPlex, Perseus** are ***easy to use*** but can handle only ***small*** complexes
- The **dual** implementation in **Dionysus** is suited to ***medium*** size complexes
- **GUDHI** and **DIPHA** are the ***most powerful*** libraries currently available
  - Can handle well ***large*** complexes

arxiv 2015, N. Otter, M. A. Porter, U. Tillmann, P. Grindrod, H. A. Harrington

Questions?

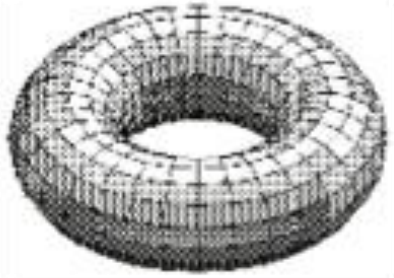
Part II

Part II Outline

- Motivating ideas
- TDA – Overview and Methodology
- 2 Examples of applications
  1. c-MYB+ Subtype of Breast Cancer
  2. TDA for Fragile X Syndrome
- Available software
- Questions

Motivation

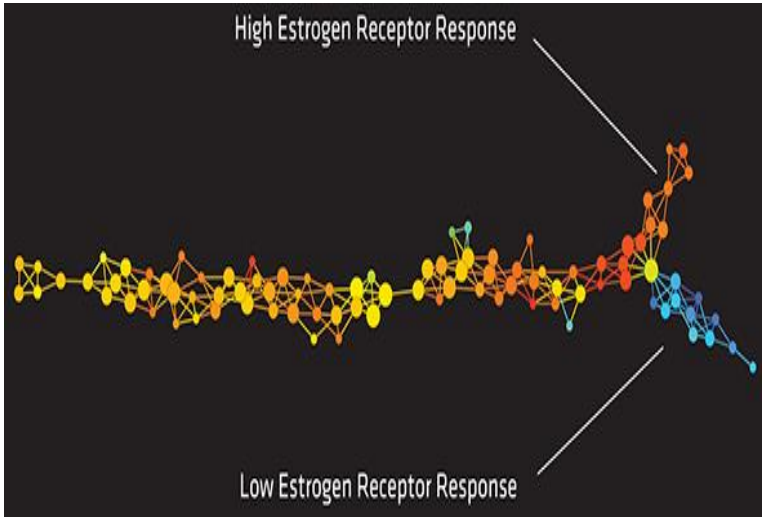
- Topological analysis can do more than just looking for loops
- TDA is another blend of topological analysis
- Produce insightful visualization
- Guides generation of hypotheses



$\beta_0 = 1$

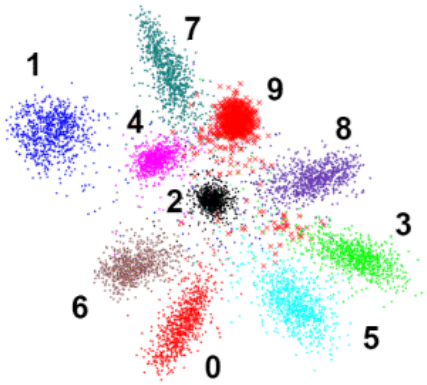
$\beta_1 = 2$

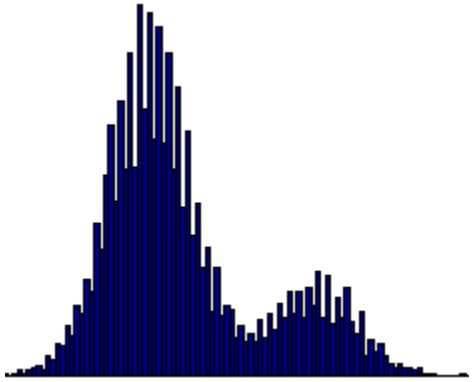
$\beta_2 = 1$

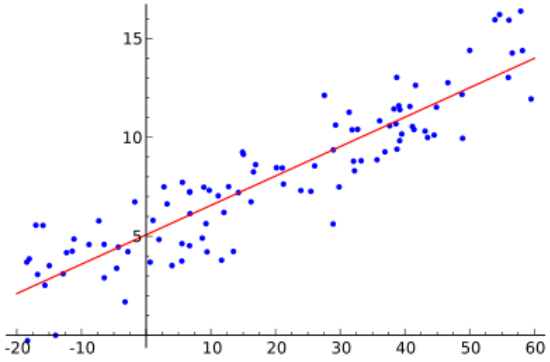


Motivation

- Shape characteristics of data have always been important to data analysis







Motivation

- An important subset of data is genomic data

Example: clustering of breast cancer microarray data; 5000 genes (columns), 98 tumor samples (rows)

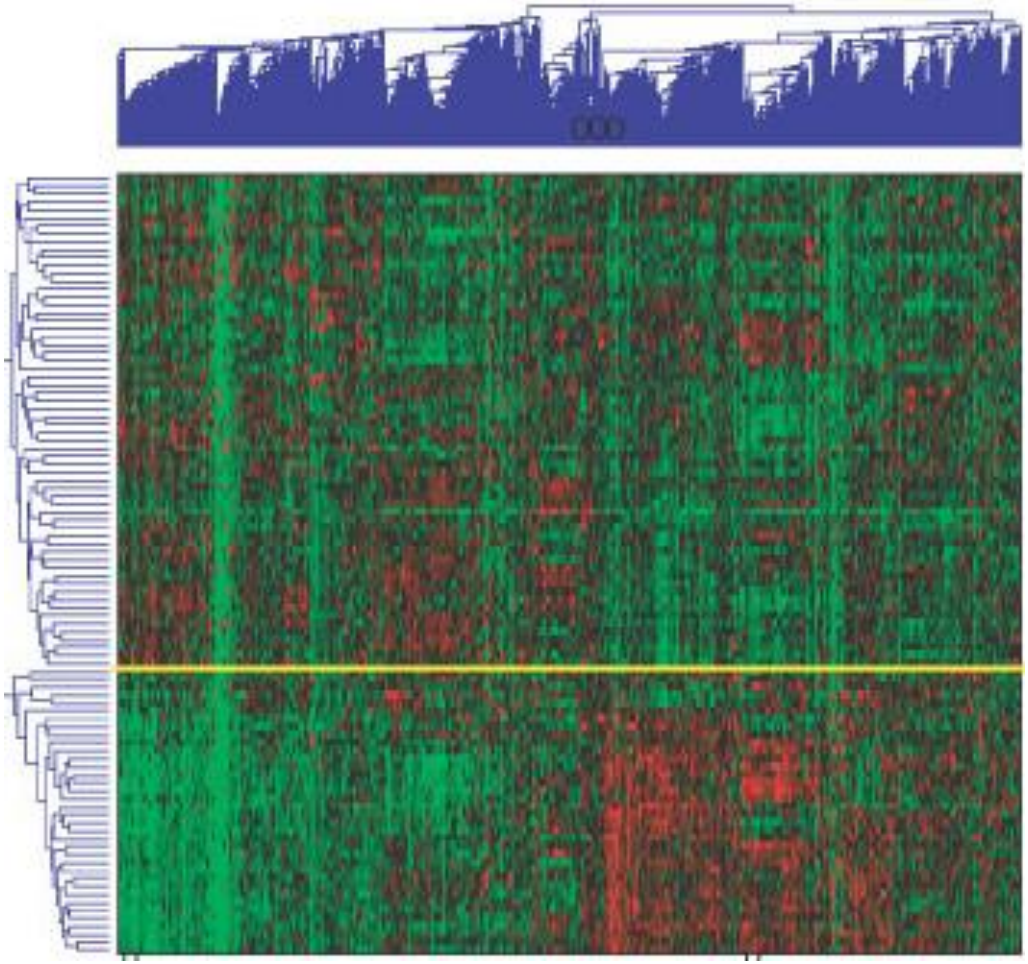
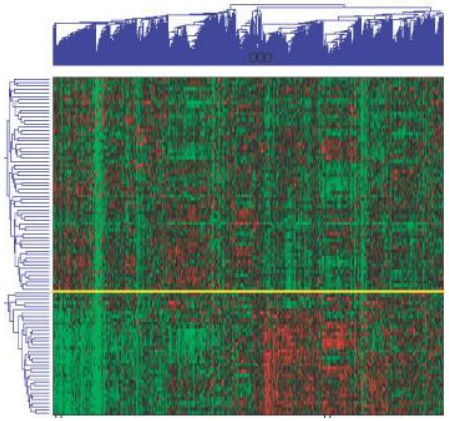


Figure from: van 't Veer, et al., Gene expression profiling predicts clinical outcome of breast cancer , Nature 2002

Motivation

- Characteristics of data:
  1. High-dimensional
  2. Complex interaction
  3. Conventional visualization gives poor hints

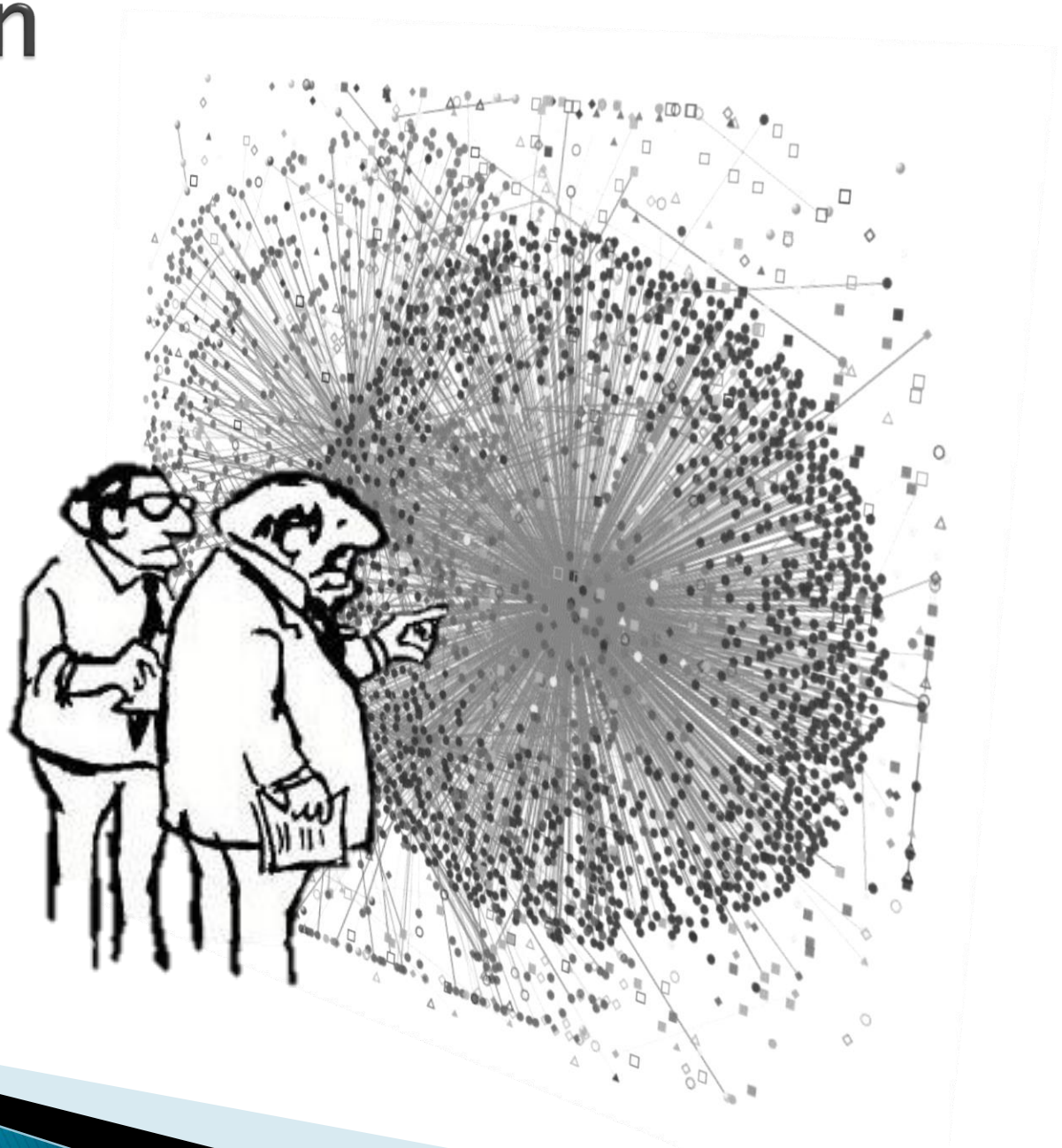


Result:

Motivation

Loss of intuition

-So, what does all this mean?



Motivation

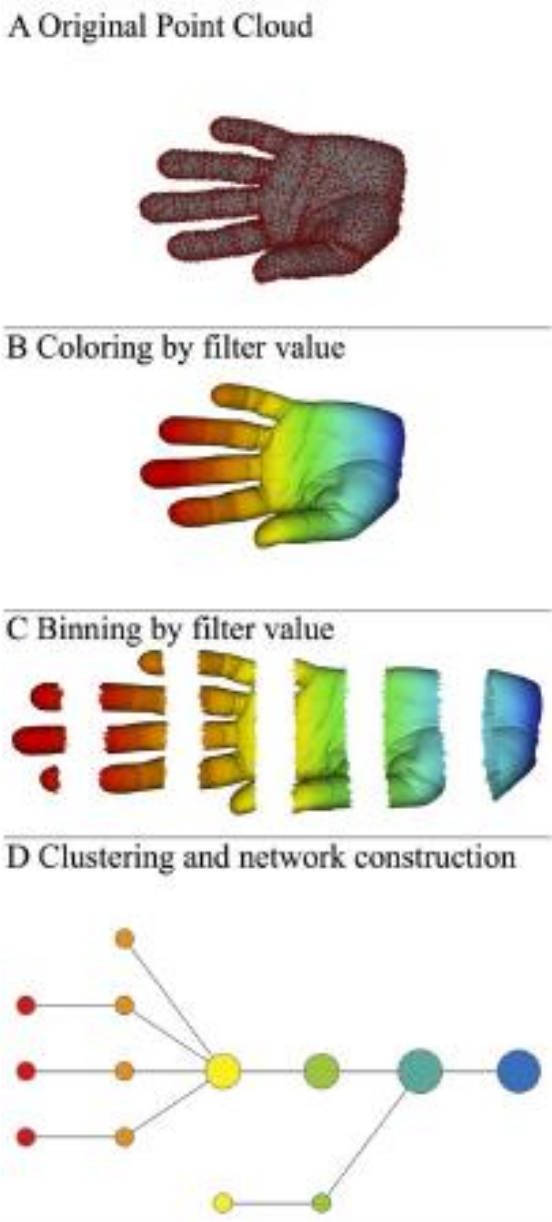
- Yet understanding data is crucially important (hypothesis generation, for example)
- How to address complexity of large-scale data?

Nature Scientific Reports 2013 – "Extracting insights from the shape of complex data using topology"

Lum, P. Y., G. Singh, A. Lehman, T. Ishkanov, Mikael Vejdemo-Johansson, M. Alagappan, J. Carlsson, and G. Carlsson.

TDA: Overview

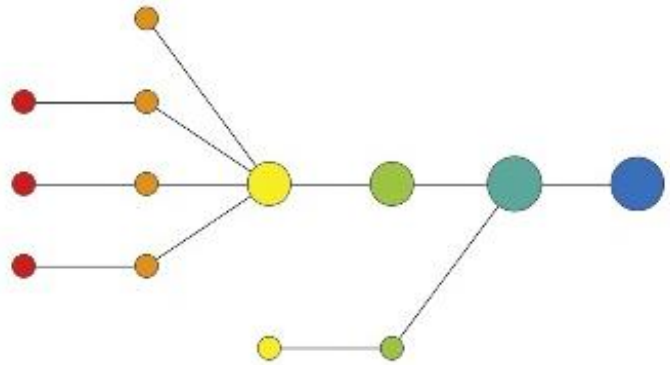
1. Begin with point cloud data
2. Assign numerical values to each point in the point cloud (color shows filter value)
   - Think of filter function as lenses you choose to look at your data. Choose your filters wise (later)
3. Separate data into overlapping bins (by filter value)
4. Cluster data points in each partition



Nature Scientific Reports 2013, Lum, et al.

TDA: Overview

- Nodes represent clusters of data points
- Edges – overlapping data points across clusters



- Data compression – only 13 nodes and 12 edges
- The technique rests on finding good filter functions

Nature Scientific Reports 2013, Lum, et al.

TDA : Filters

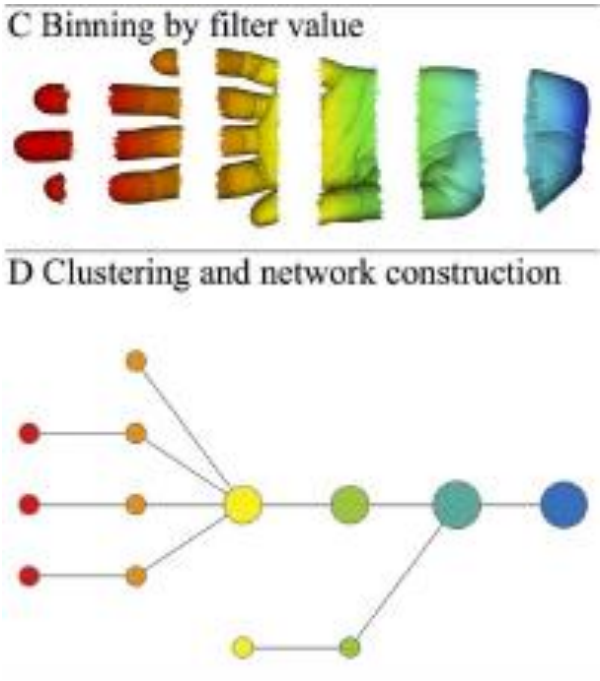
- Filters are mathematical functions through which you see the data points
- Some examples of filter functions:

Geometry	Statistics	ML	Data Driven
Euclidean distance	Max/Min	PCA	Died/Lived
Density estimator	Average	SVD	Age
Correlation distance	Variance	…	Dates
Various L-norms	…	…	…
…	…	…	…

- In general, can have more filters than 1

TDA : Clustering

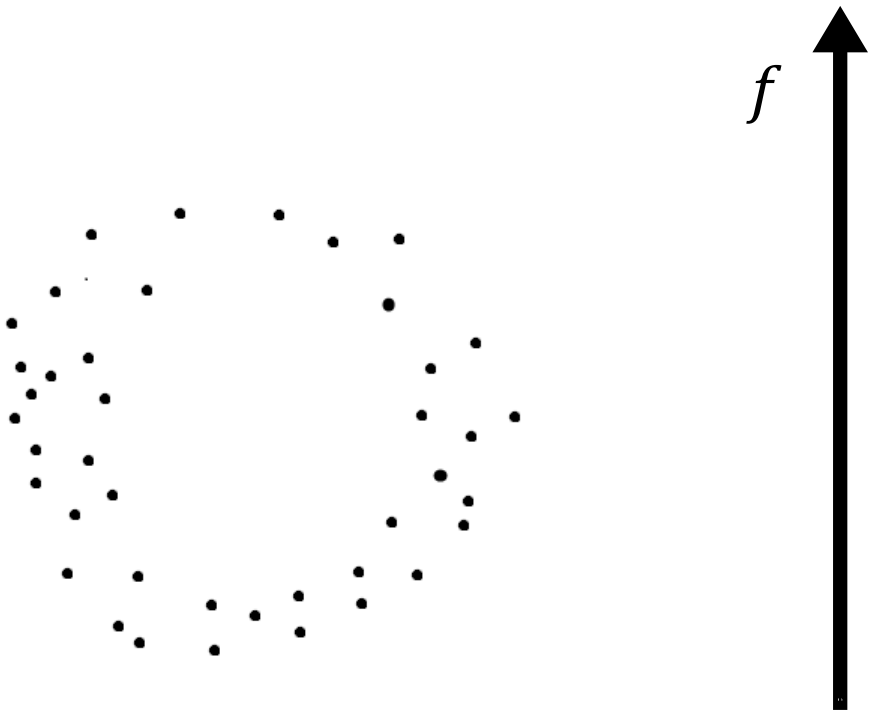
- Within each partition data points are clustered
- Ex. single linkage clustering, can use other clustering methods
- Choose metric for clustering
- Some examples – Euclidean, variance normalized Euclidean, correlation, etc.



Nature Scientific Reports 2013, Lum, et al.

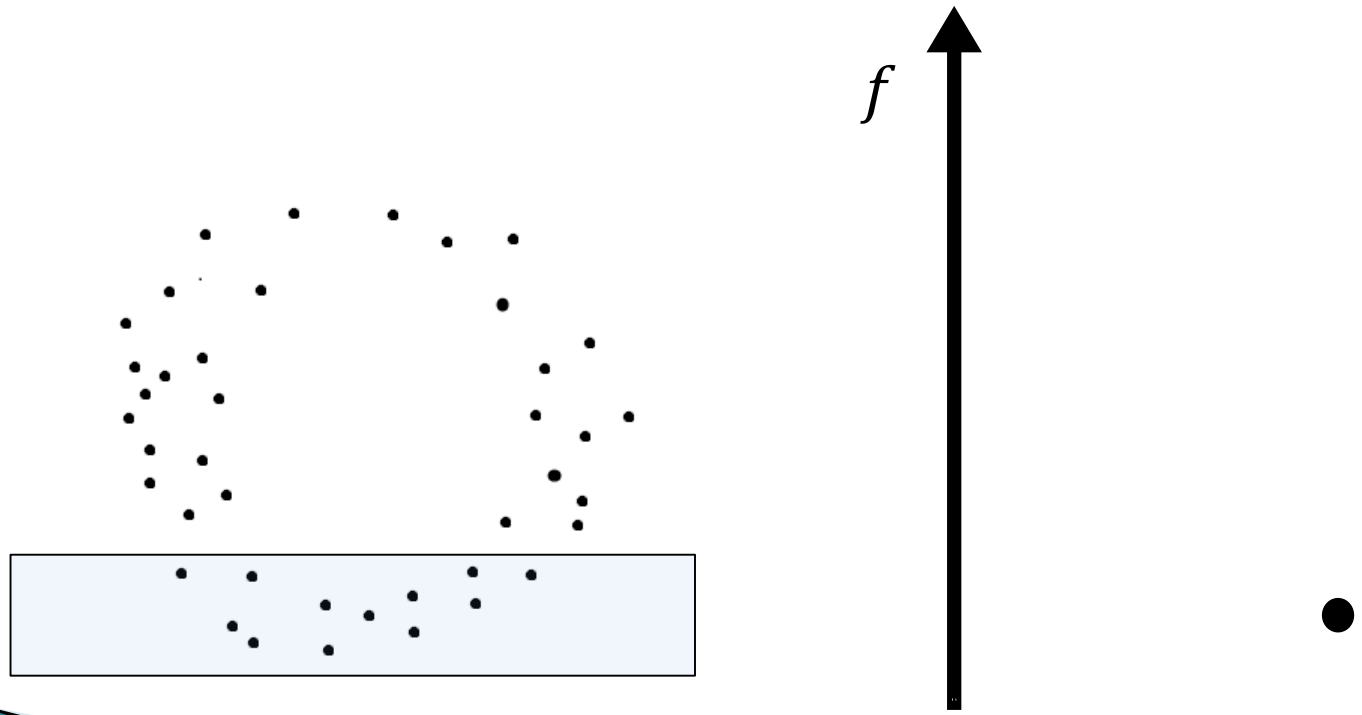
TDA : Hands-On Example

- Filter function f, binning with $N=6$ bins
- Overlap k ~30%
- Single linkage clustering, metric – Euclidean



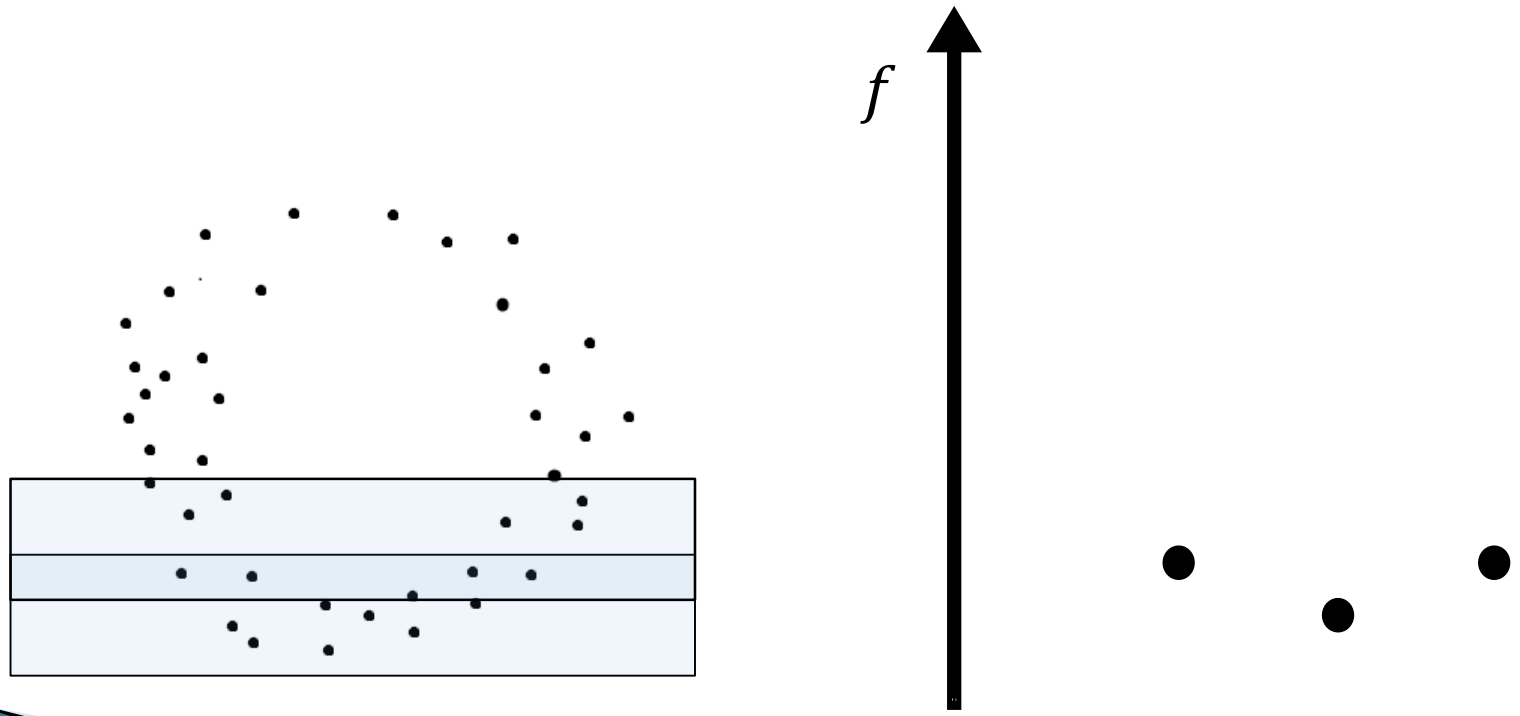
TDA : Hands-On Example

- Filter function f, binning with $N=6$ bins
- Overlap k ~30%
- Single linkage clustering, metric – Euclidean



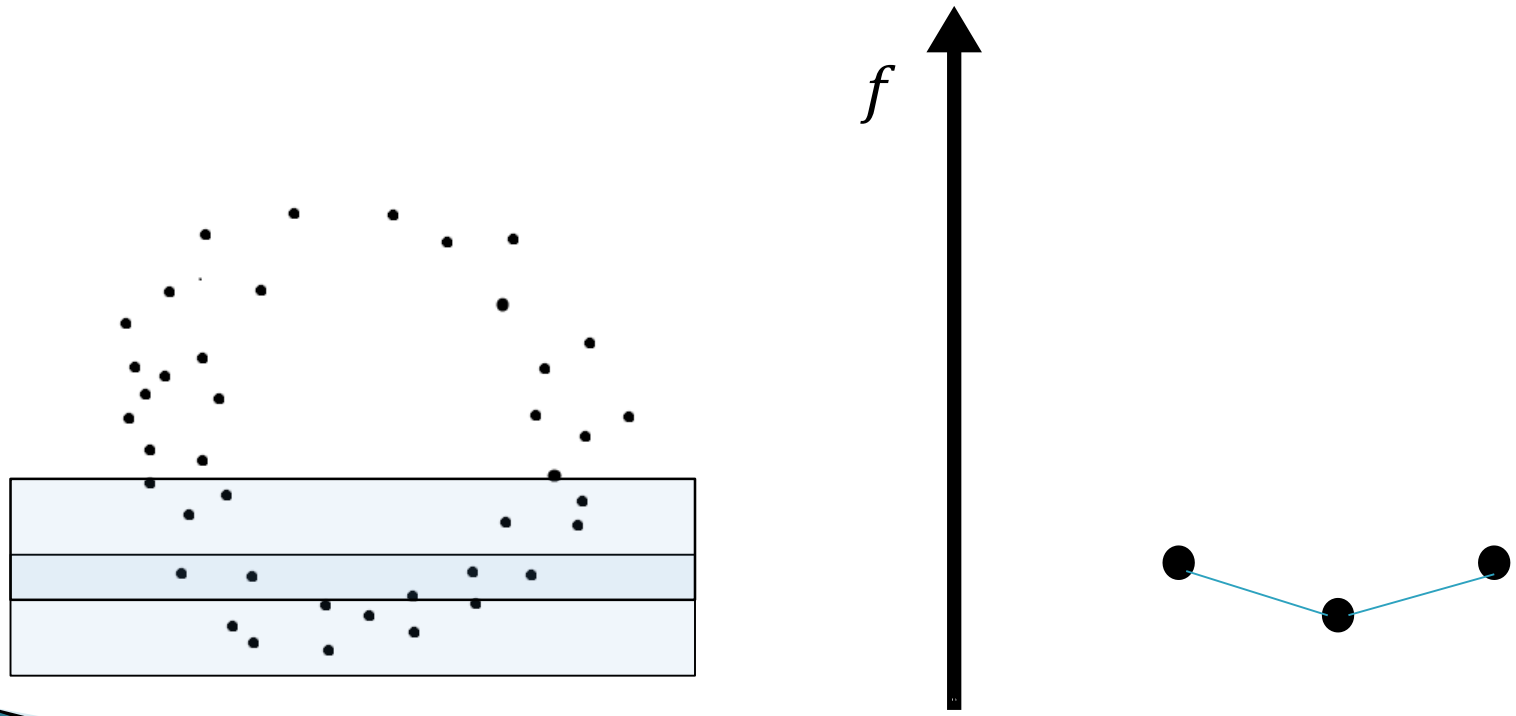
TDA : Hands-On Example

- Filter function f, binning with $N=6$ bins
- Overlap k ~30%
- Single linkage clustering, metric – Euclidean



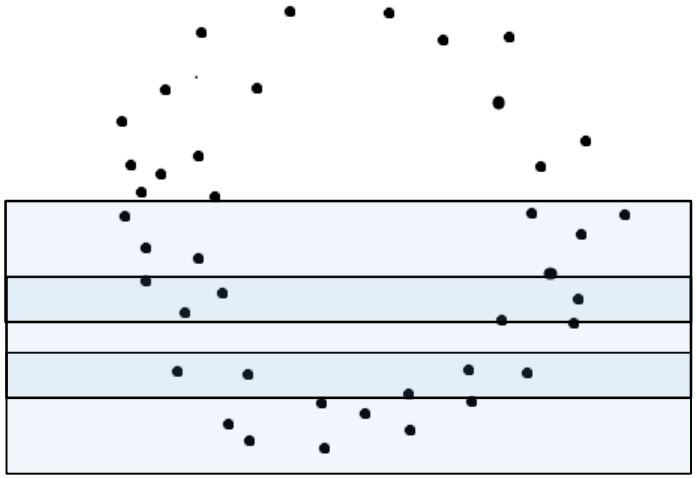
TDA : Hands-On Example

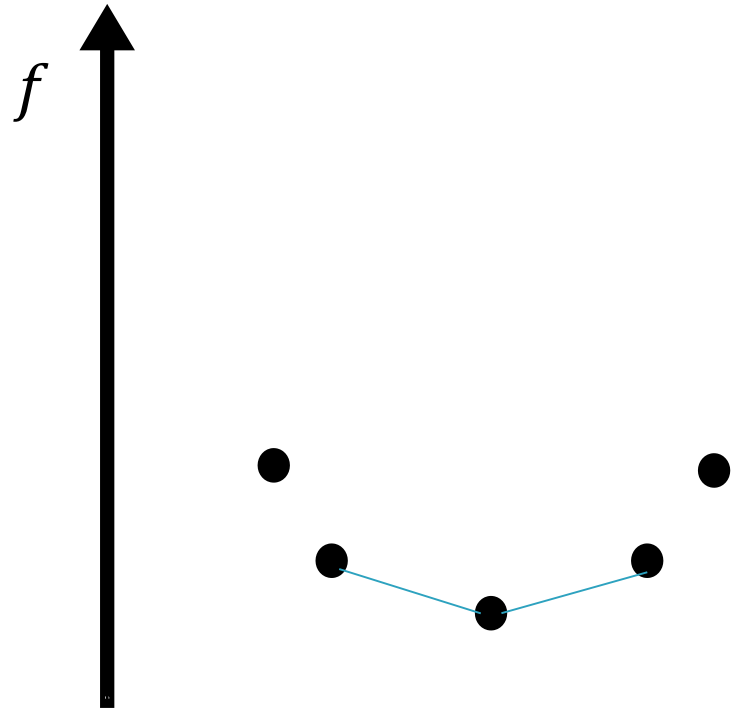
- Filter function f, binning with $N=6$ bins
- Overlap k ~30%
- Single linkage clustering, metric – Euclidean



TDA : Hands-On Example

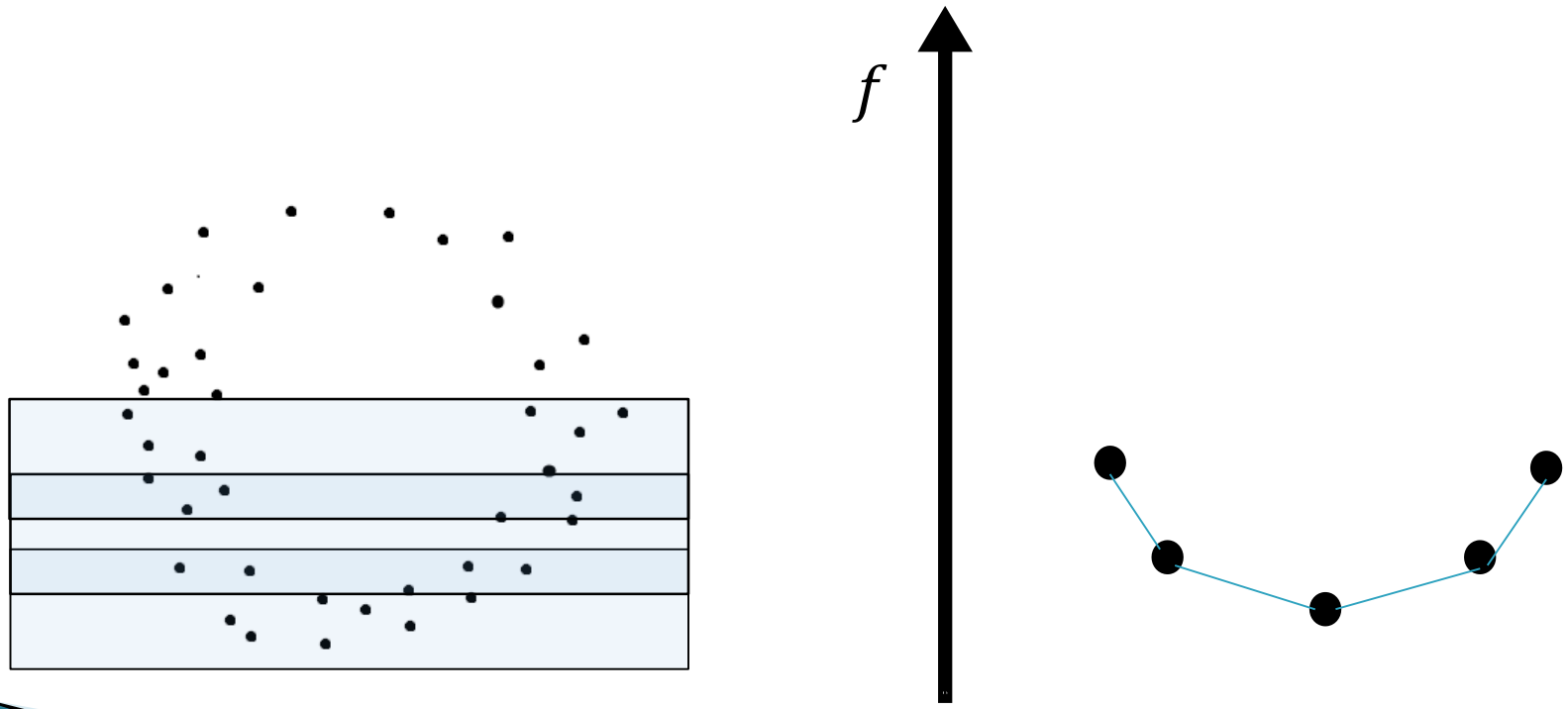
- Filter function f, binning with $N=6$ bins
- Overlap k ~30%
- Single linkage clustering, metric – Euclidean





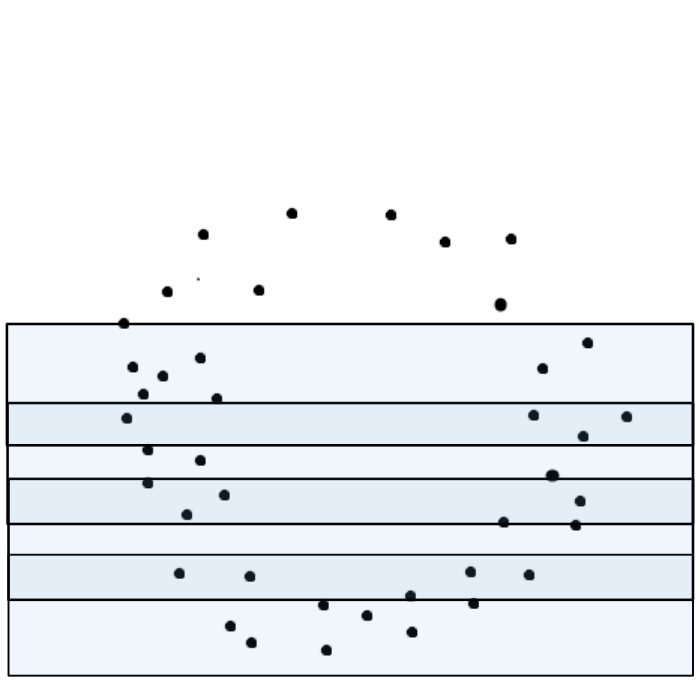
TDA : Hands-On Example

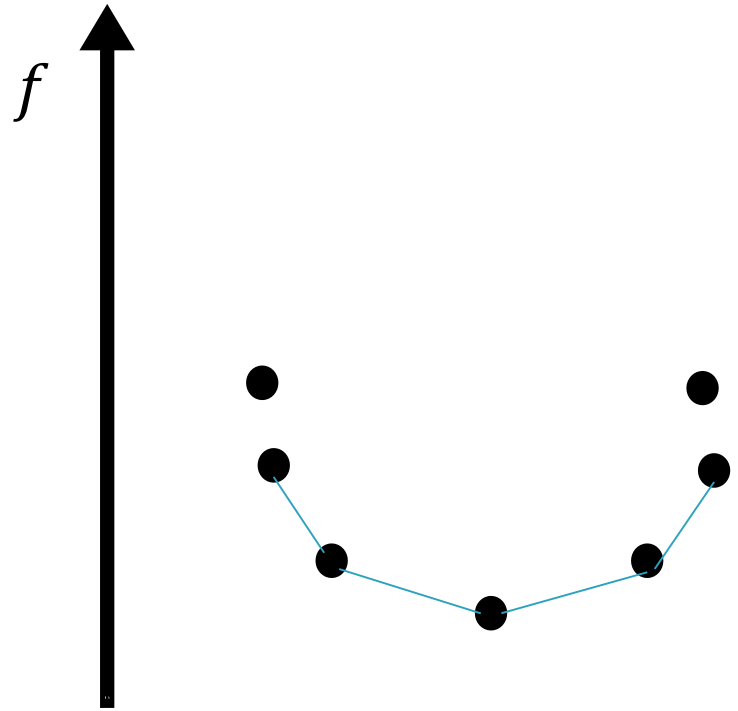
- Filter function f, binning with $N=6$ bins
- Overlap k ~30%
- Single linkage clustering, metric – Euclidean



TDA : Hands-On Example

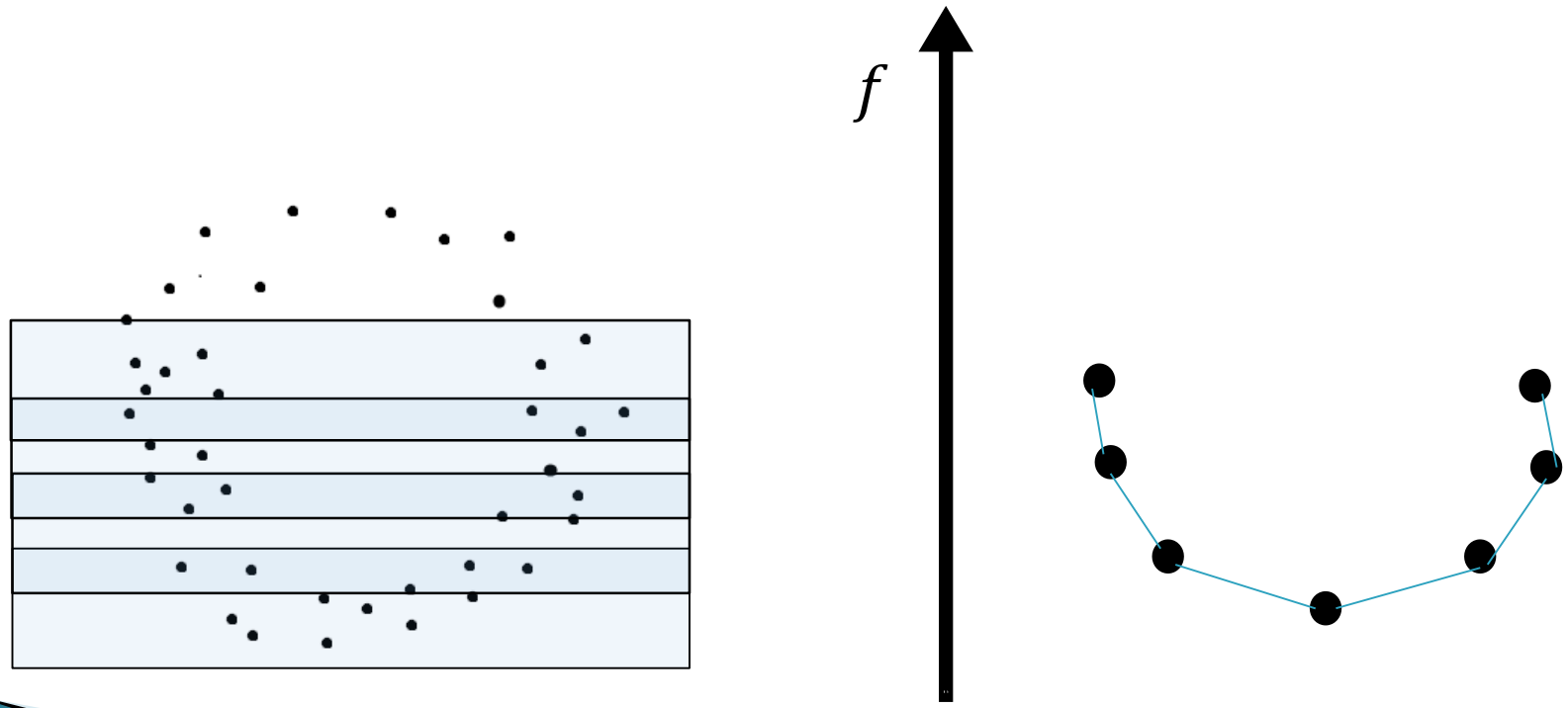
- Filter function f, binning with $N=6$ bins
- Overlap k ~30%
- Single linkage clustering, metric – Euclidean





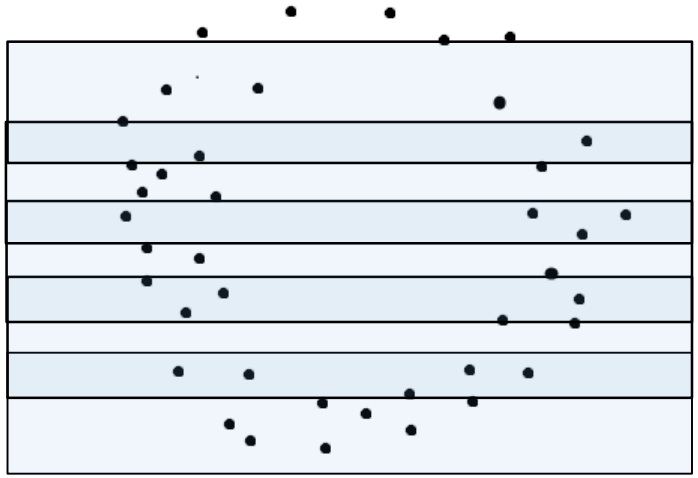
TDA : Hands-On Example

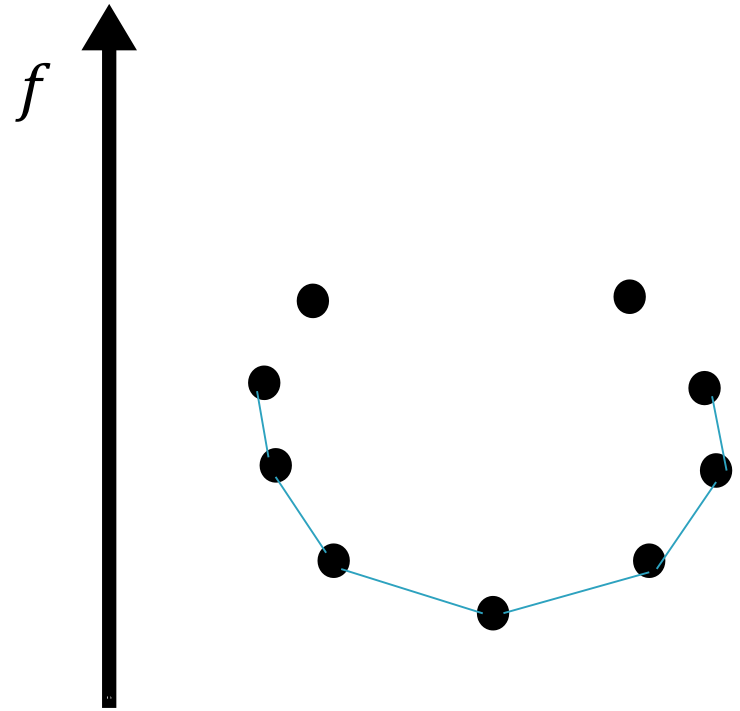
- Filter function f, binning with $N=6$ bins
- Overlap k ~30%
- Single linkage clustering, metric – Euclidean



TDA : Hands-On Example

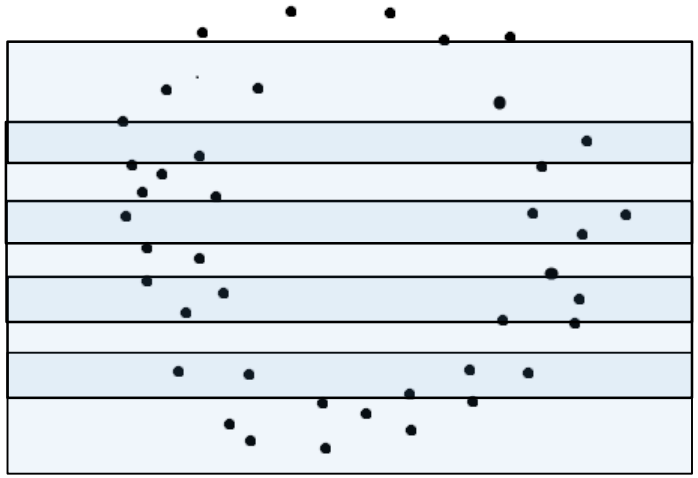
- Filter function f, binning with $N=6$ bins
- Overlap k ~30%
- Single linkage clustering, metric – Euclidean

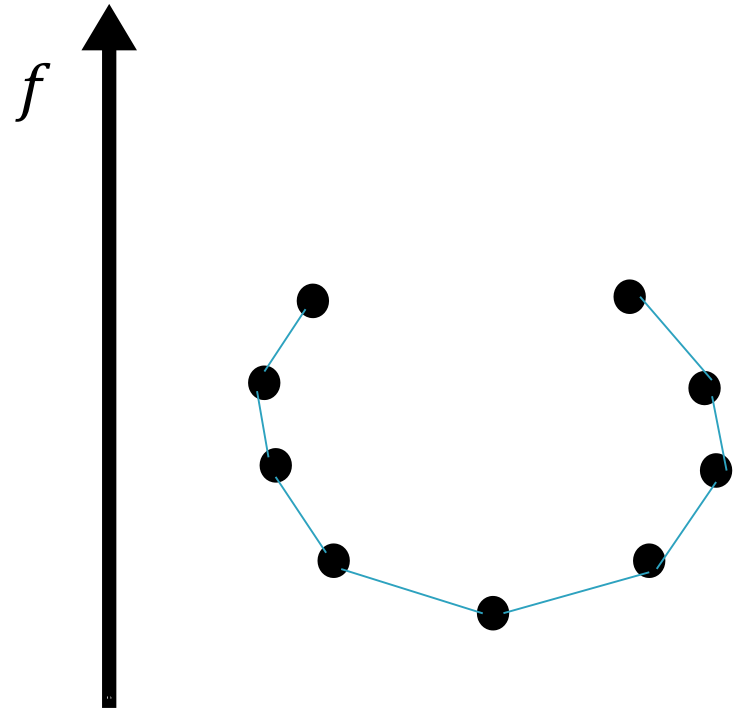




TDA : Hands-On Example

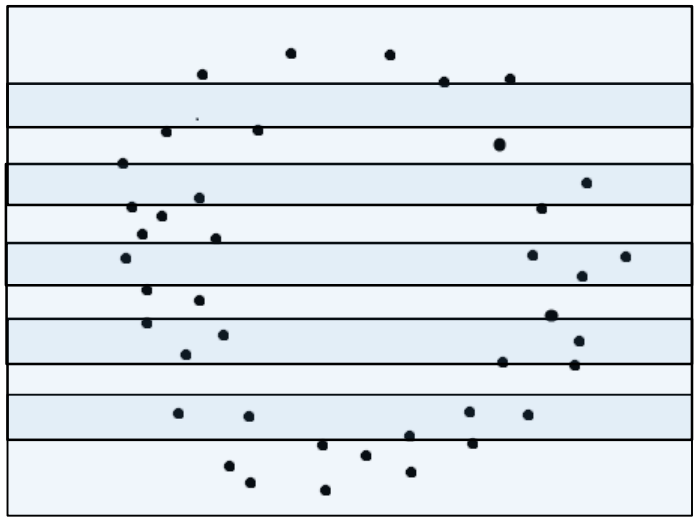
- Filter function f, binning with $N=6$ bins
- Overlap k ~30%
- Single linkage clustering, metric – Euclidean

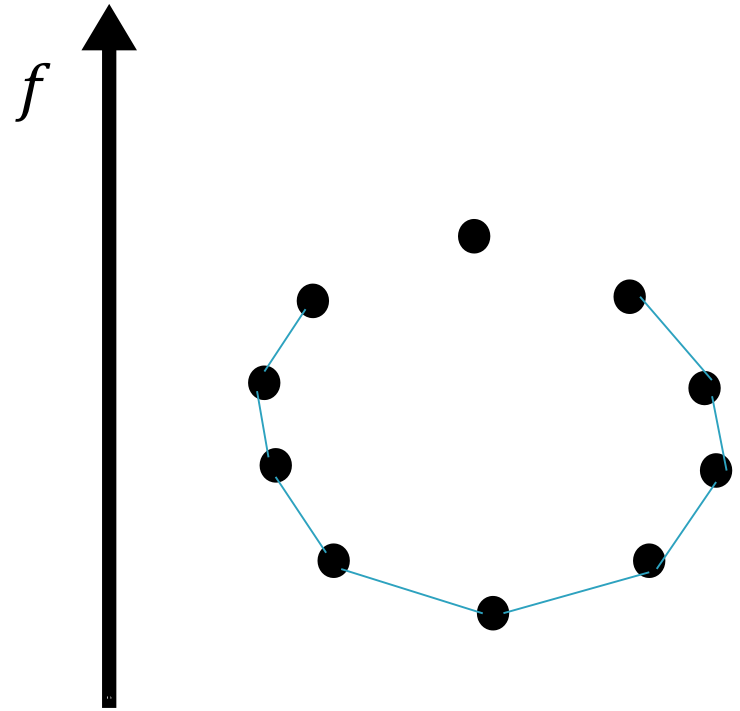




TDA : Hands-On Example

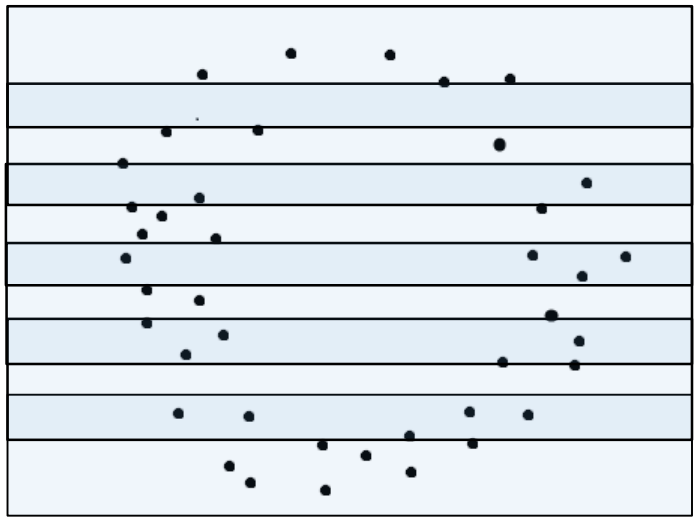
- Filter function f, binning with $N=6$ bins
- Overlap k ~30%
- Single linkage clustering, metric – Euclidean

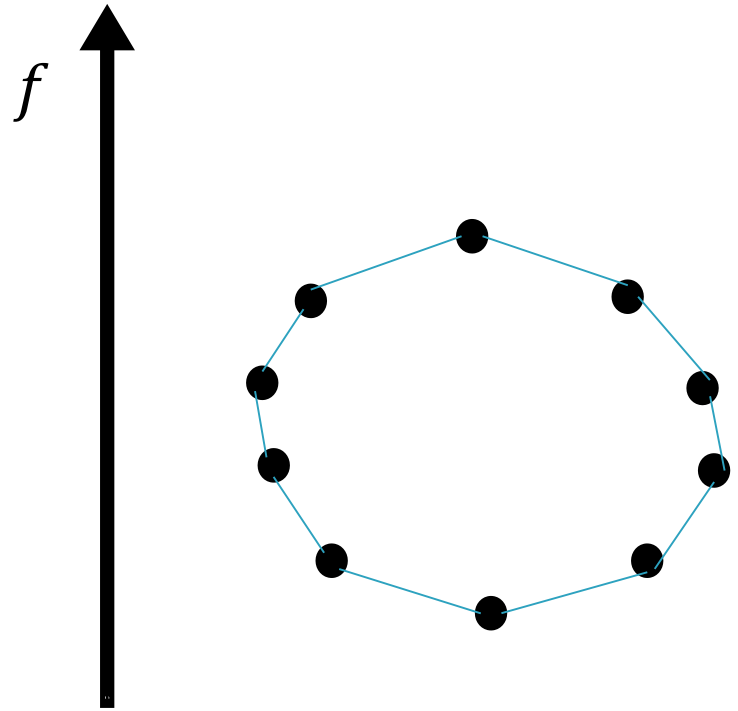




TDA : Hands-On Example

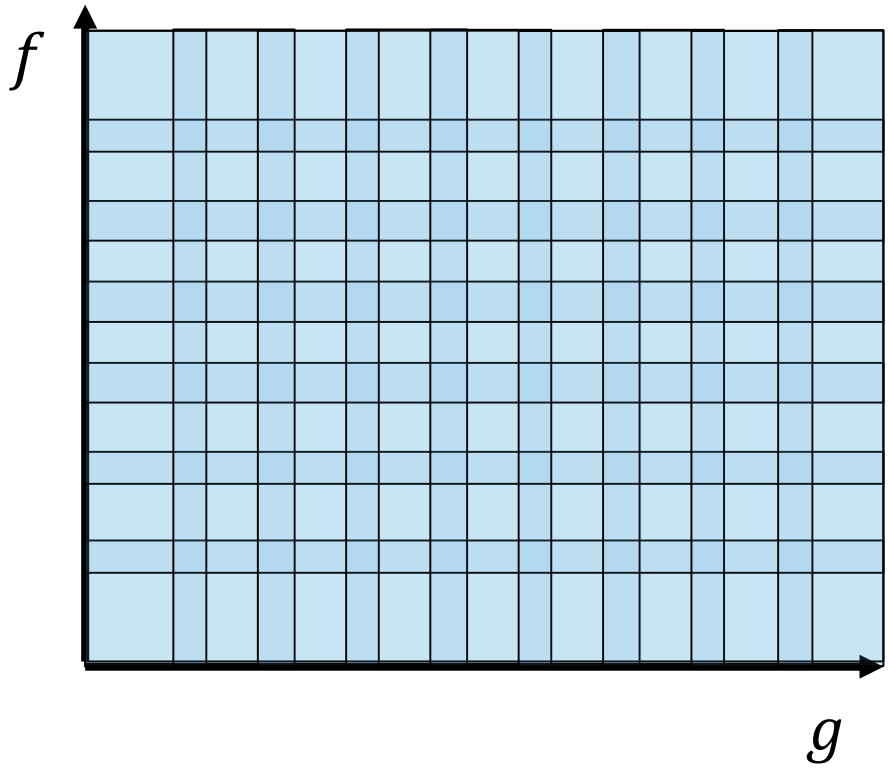
- Filter function f, binning with $N=6$ bins
- Overlap k ~30%
- Single linkage clustering, metric – Euclidean





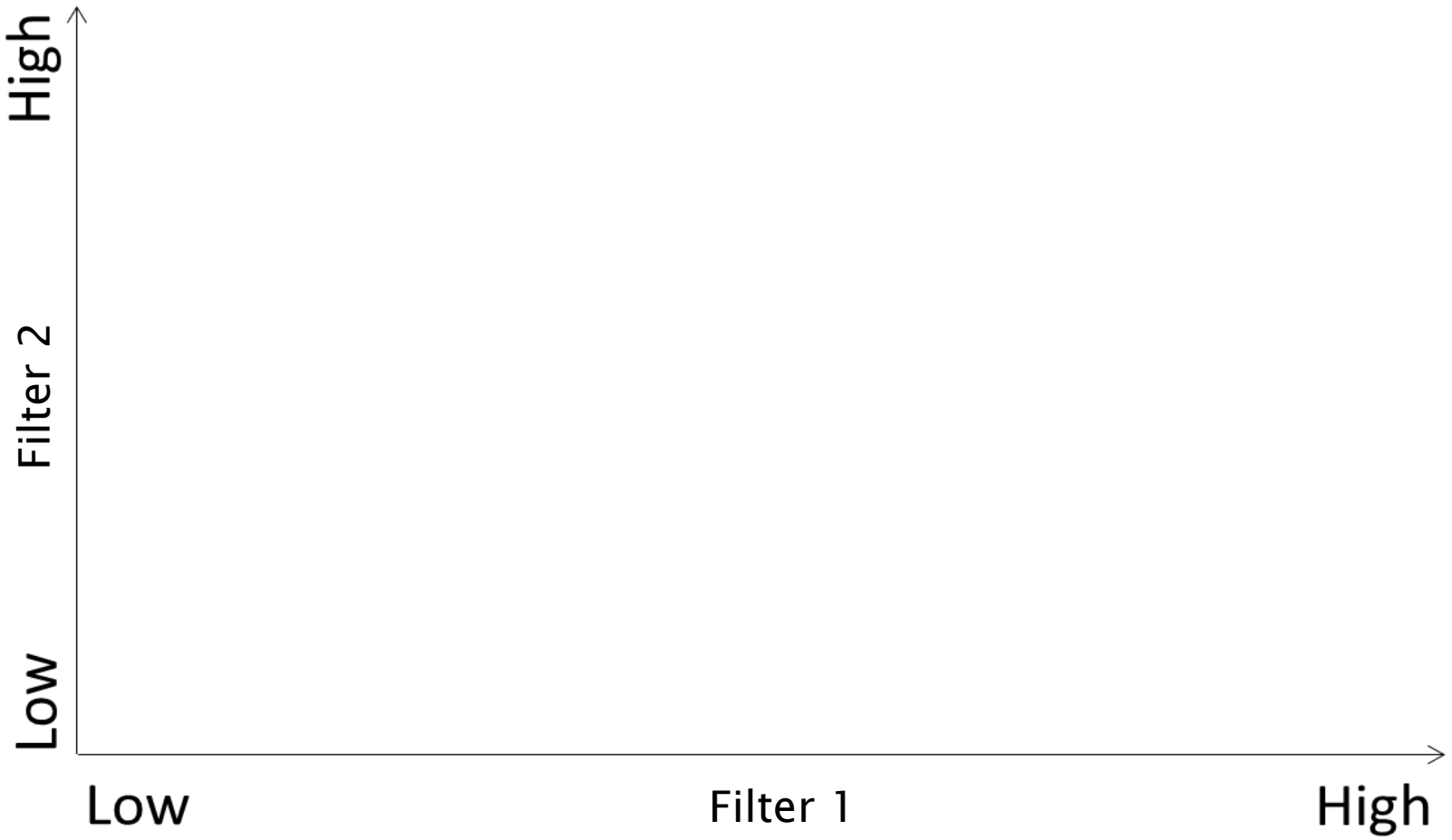
TDA : More Details

- In general, can have more filters than 1



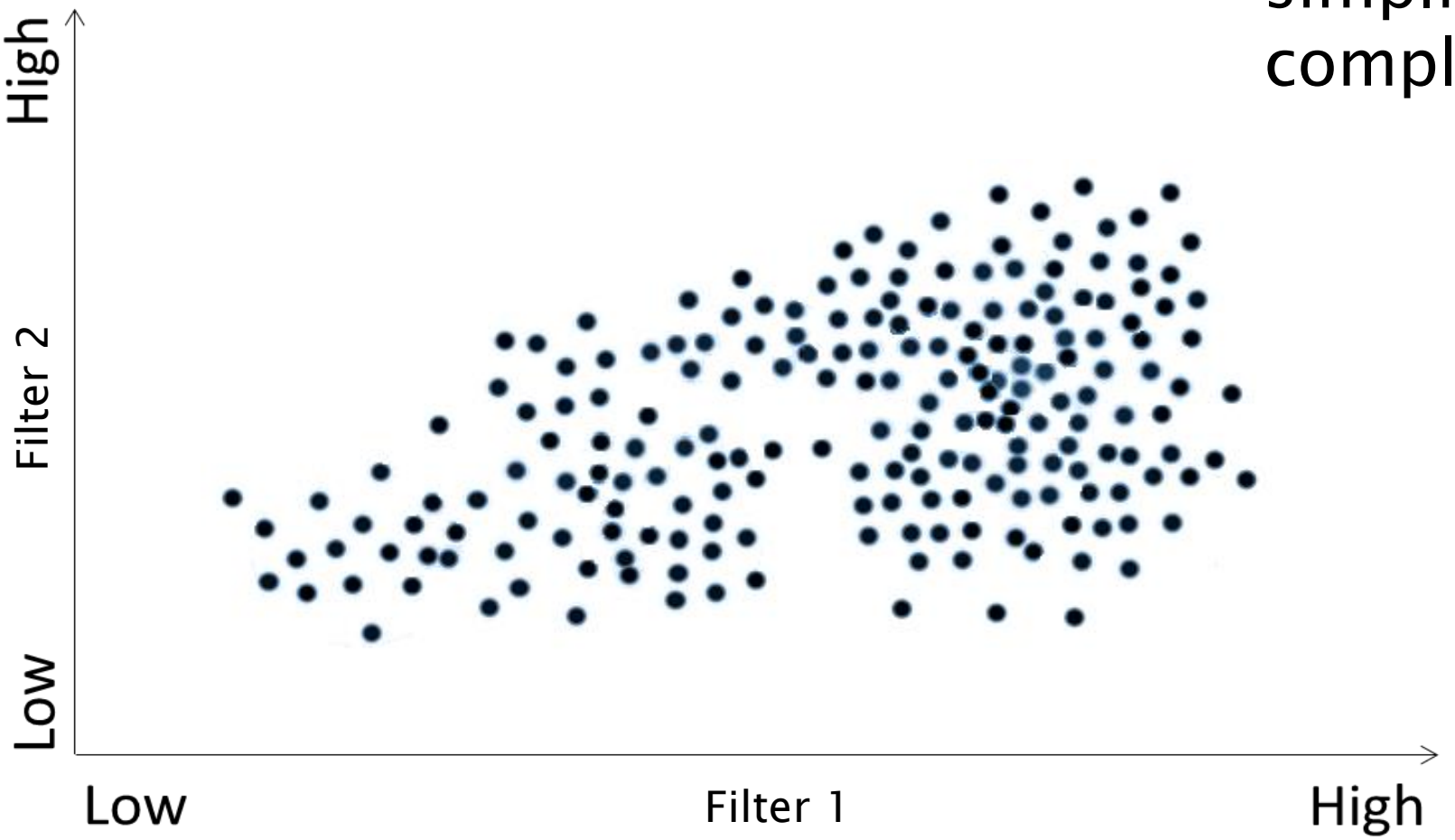
TDA : More Details

- Also, no need to restraint to nodes/edges



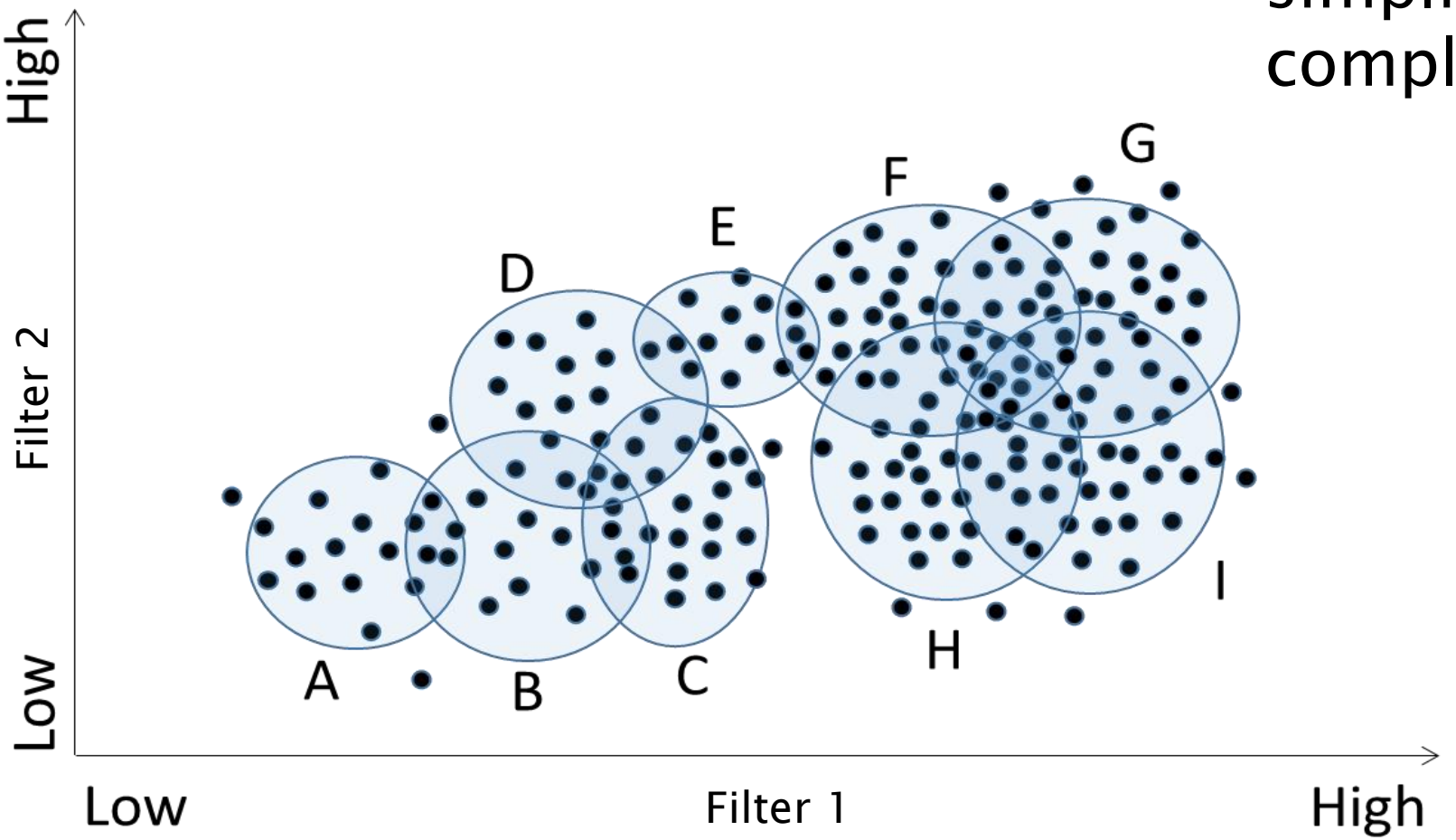
TDA : More Details

- Construct higher simplicial complexes



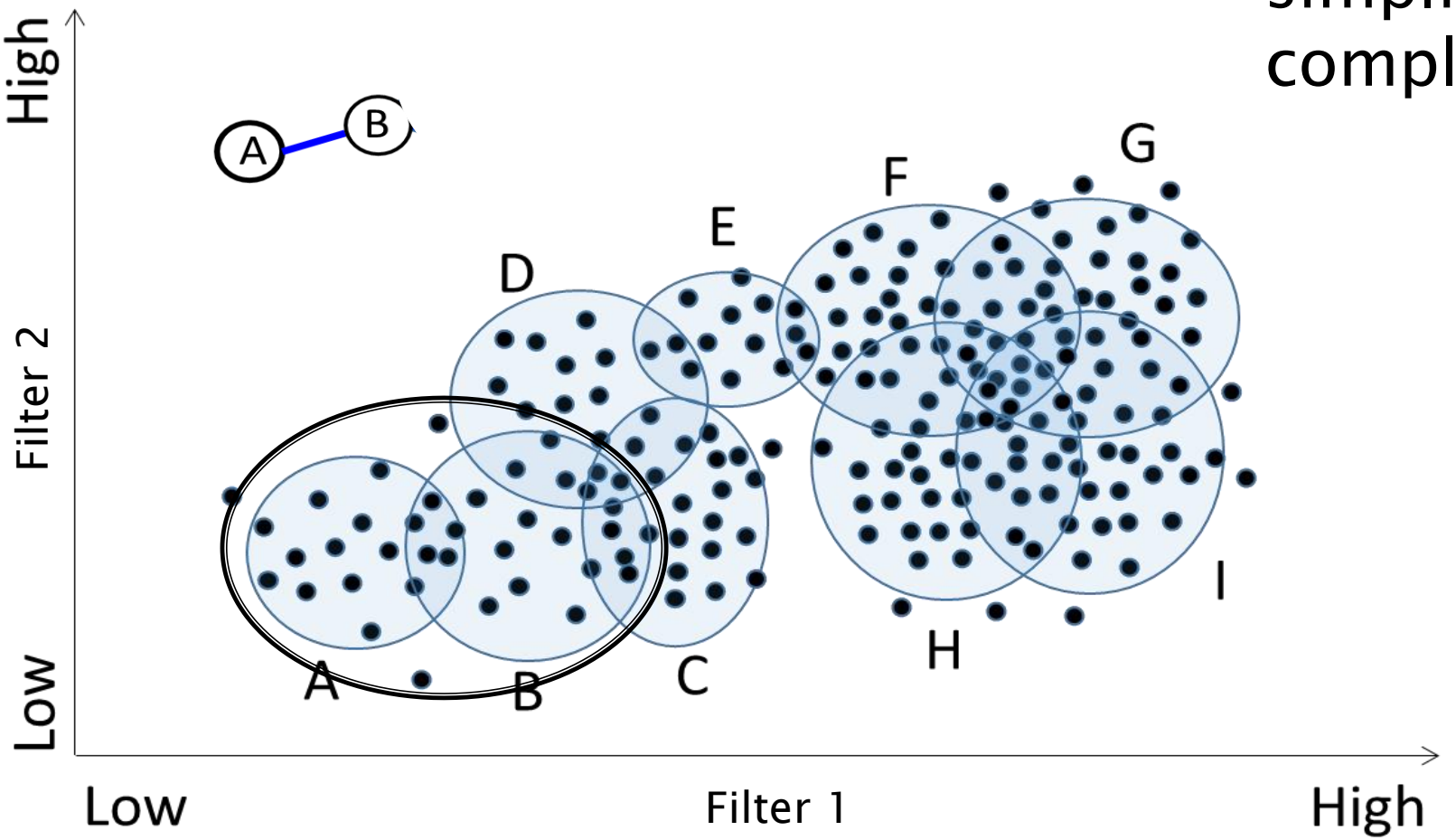
TDA : More Details

- Construct higher simplicial complexes



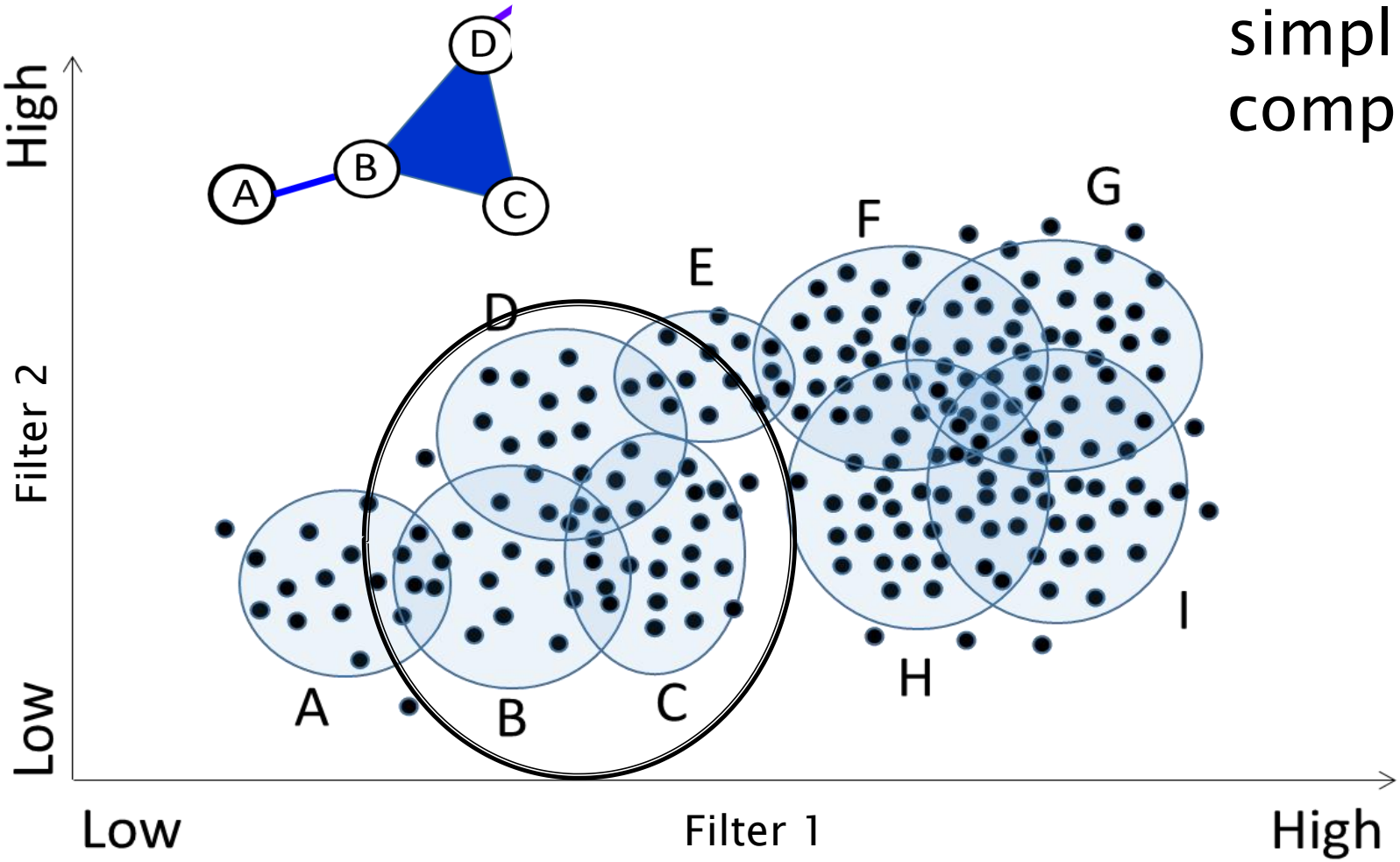
TDA : More Details

- Construct higher simplicial complexes



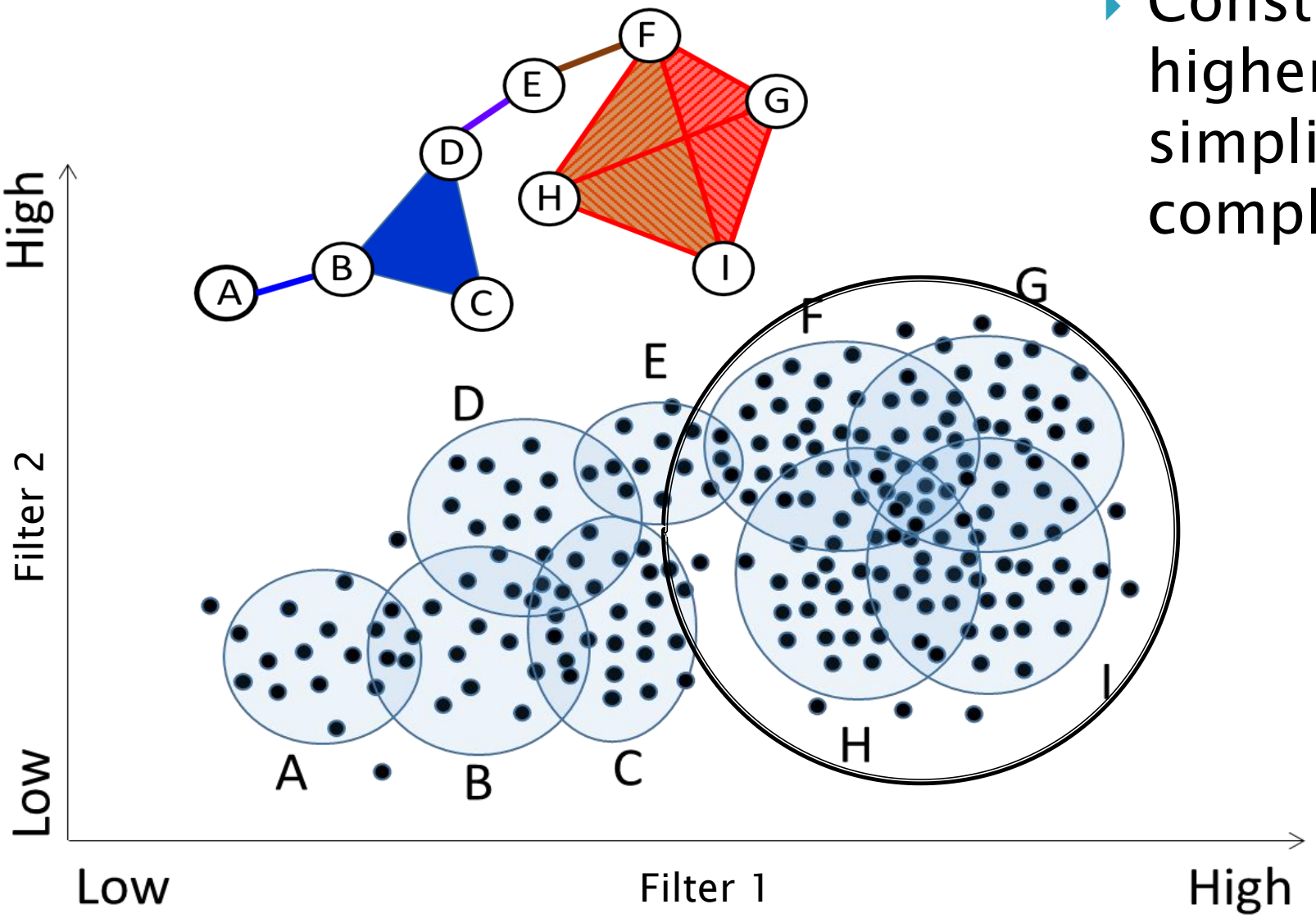
TDA : More Details

- Construct higher simplicial complexes



TDA : More Details

- Construct higher simplicial complexes

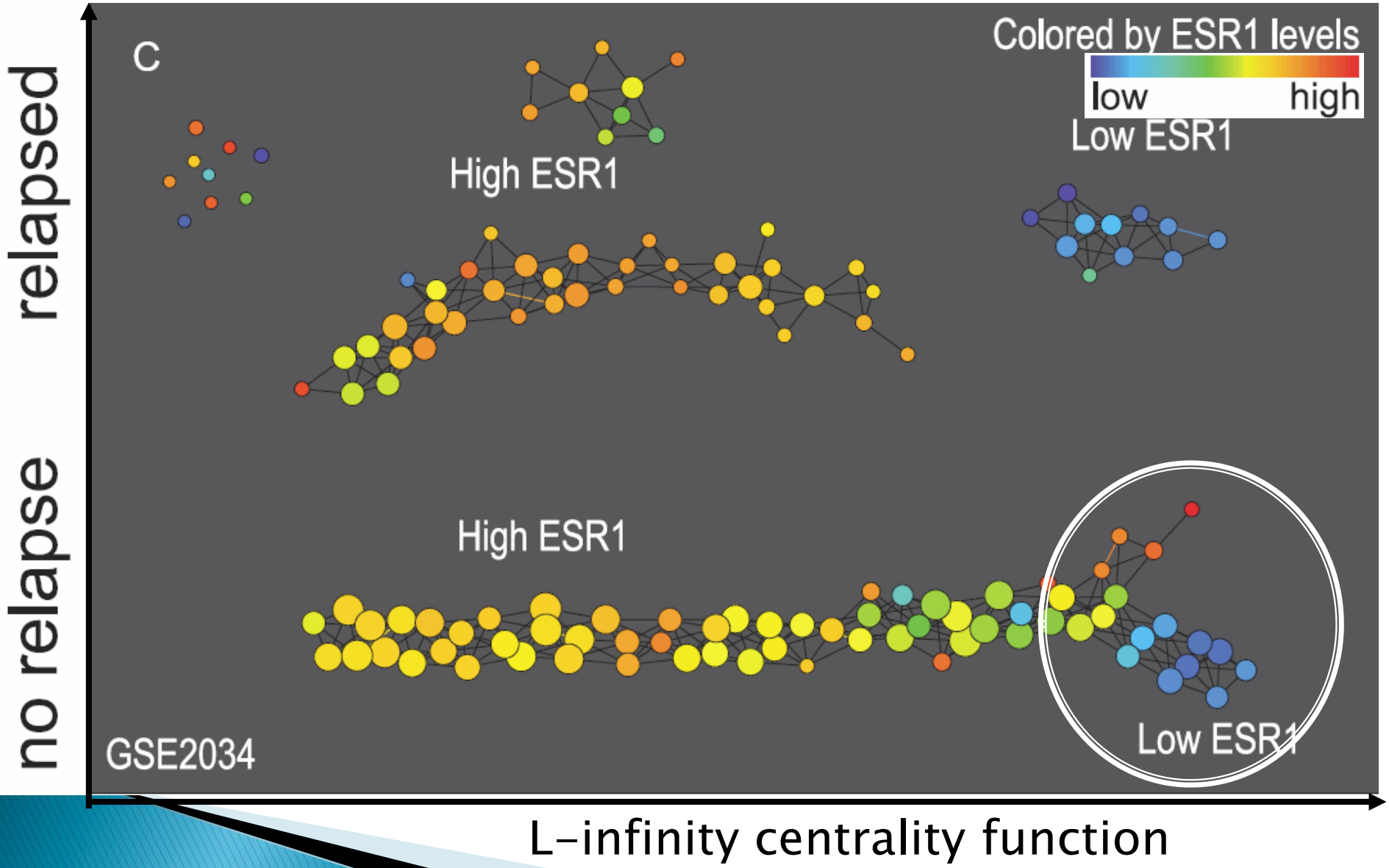


TDA Visualization Examples

- TDA compresses data while highlighting critical aspects of it

Nature Scientific Reports 2013, Lum, et al.

GSE2034 – 286 tumors, 17 819 genes

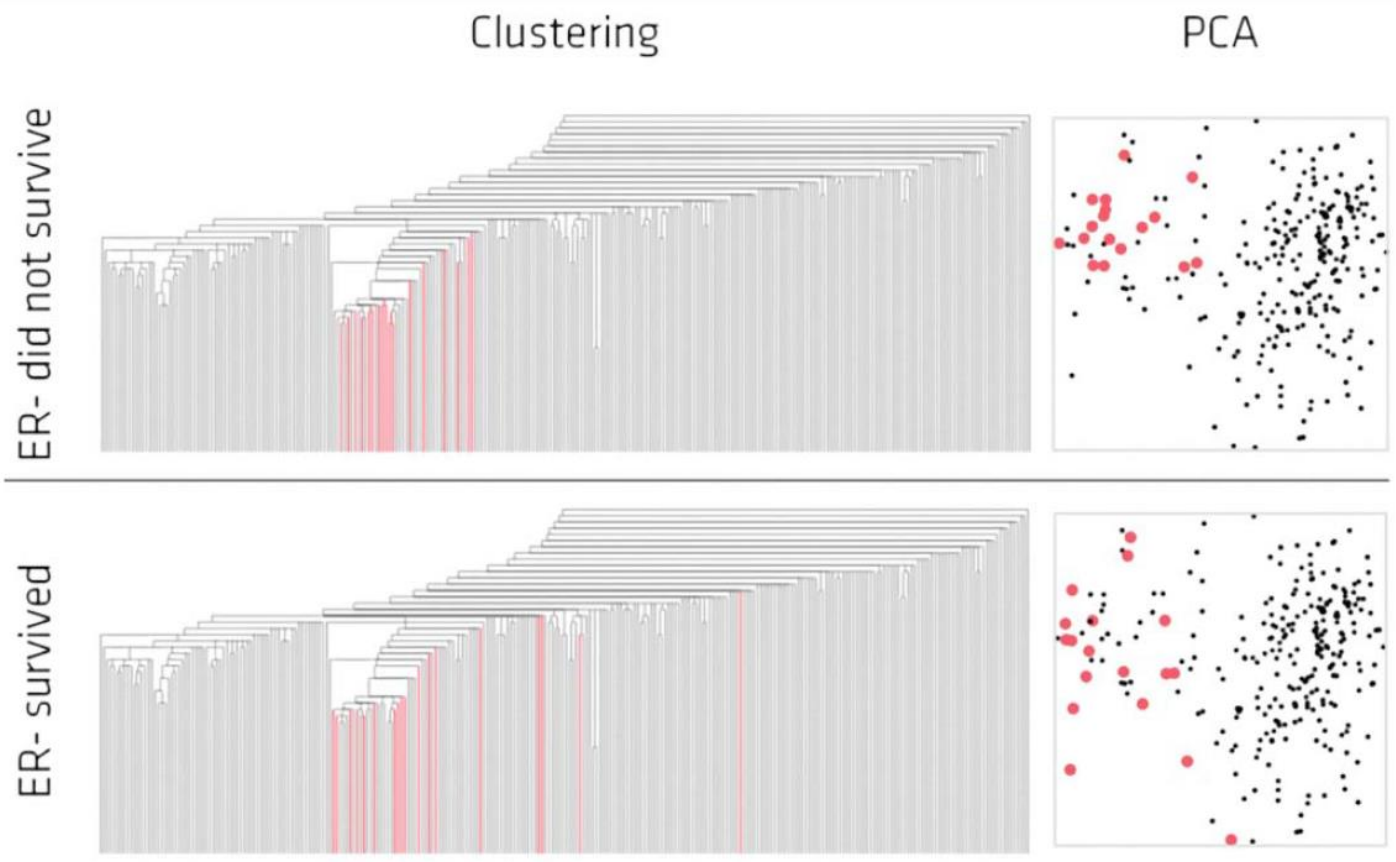


Nature Scientific Reports 2013, Lum, et al.

TDA of Breast Cancer

- In contrast, traditional analysis with PCA and clustering – no clear delineation

Highlighted in red are low ESR1 patients

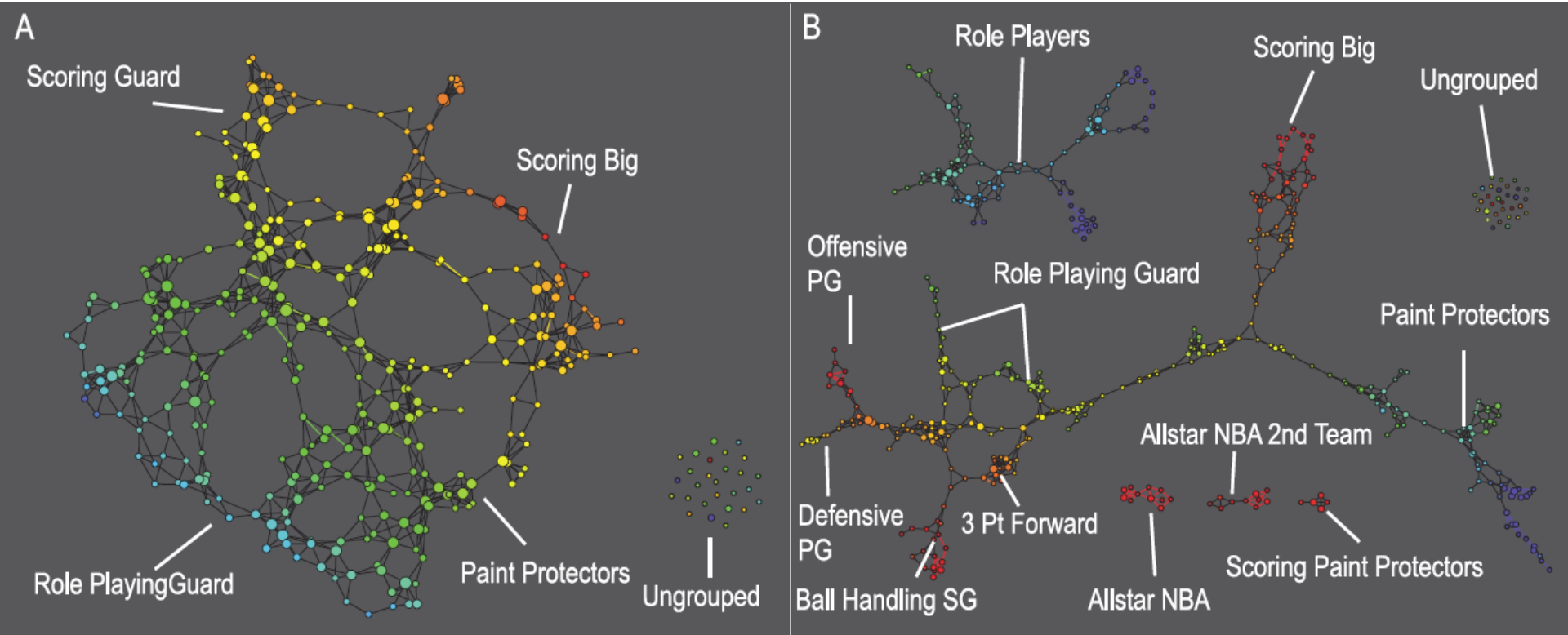


Nature Scientific Reports 2013, Lum, et al.

TDA Visualization Examples

- Sensitive to both large and small scale patterns
  - 452 NBA players, metric: variance normalized Euclidean; filters – 1st and 2nd SVD value, colored by score

lower resolution – 20 bins higher resolution – 30 bins



TDA: Summary

- Combines the best features of existing methodologies such as PCA or clustering
- Provides a compressed geometric representation of complex data sets
- Allows visual inspection at different resolution levels

Questions?

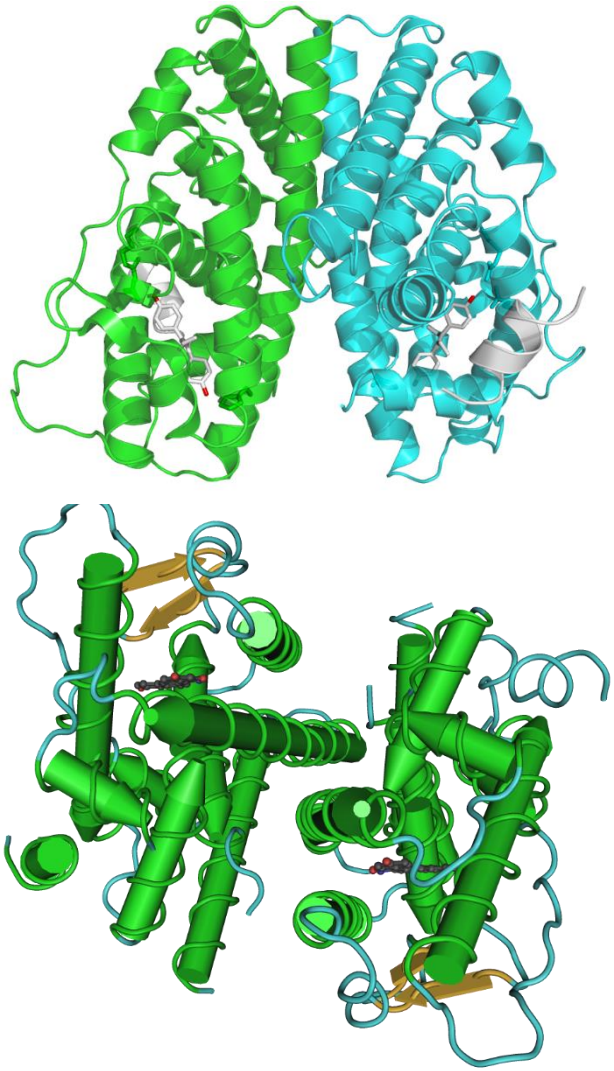
TDA Applications

PNAS 2011 – “Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival”

Nicolau, Monica, Arnold J. Levine, and Gunnar Carlsson.

c-MYB$^+$ Type of Breast Cancer

- Breast cancer has multiple sub-types
- Different subtypes exist within ER$^+$
- By applying TDA, identified a new cancer type – c-MYB$^+$
- Before TDA, data preprocessing step (*DSGA*) was applied – Why? GIGO



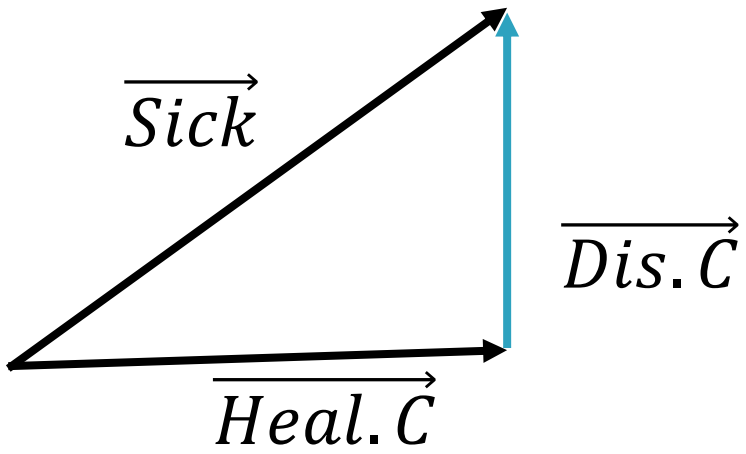
Estrogen receptors

PNAS 2011, Nicolau M, Levine AJ, Carlsson G.

DSGA transform

- Think of data as vectors:

	Gene1	Gene2	Gene3	Gene4	Gene5	Gene6	
Sick Patient	10.20	10.20	10.21	10.20	10.20	10.20	$\overrightarrow{Sick}$
Healthy Patient	10.76	10.61	10.43	10.59	10.63	10.51	$\overrightarrow{Heal.C}$



PNAS 2011, Nicolau M, Levine AJ, Carlsson G.

Visualization

- Work on DSGA-transformed data only
- Threshold data so that only that only genes with significant deviation left
- Apply filter functions on data:

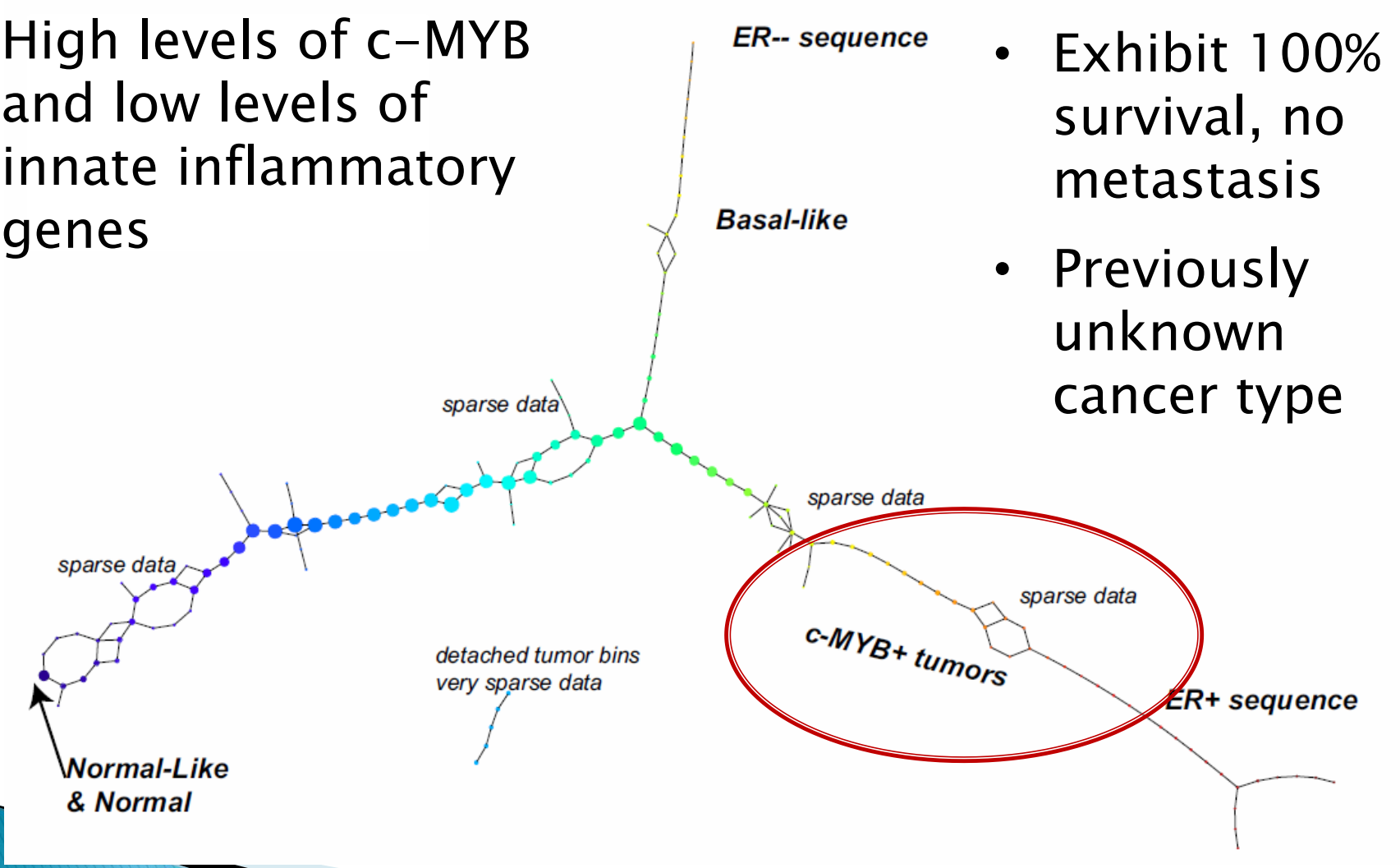
$$f_{p,k}(\vec{V}) = [\Sigma |\ g_r|^p]^{k/p}$$

  - Observe that for $k = 1$ and $p = 2$, it computes the standard Euclidean distance

PNAS 2011, Nicolau M, Levine AJ, Carlsson G.

c-MYB+ Type of Breast Cancer

- High levels of c-MYB and low levels of innate inflammatory genes

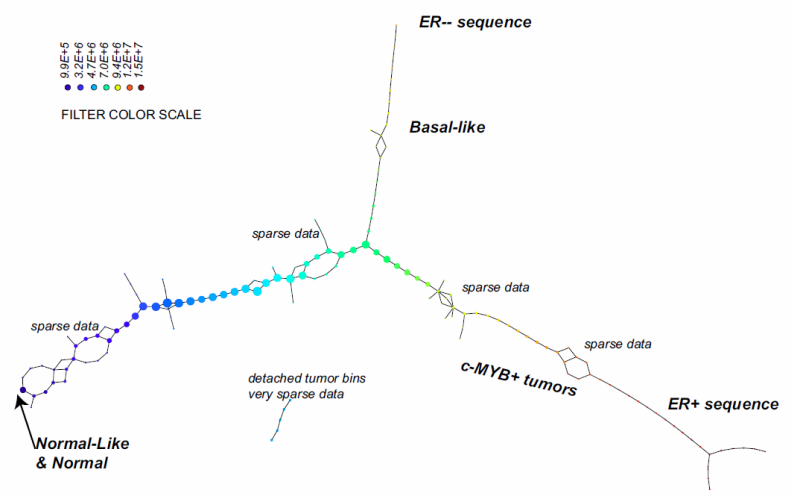


- Exhibit 100% survival, no metastasis
- Previously unknown cancer type

PNAS 2011, Nicolau M, Levine AJ, Carlsson G.

c-MYB^{+} Type of Breast Cancer

- No supervised step beyond distinction between tumor and healthy patients
- The group has statistically significant molecular signature
- Highlights coherent biology invisible to cluster methods



PNAS 2011, Nicolau M, Levine AJ, Carlsson G.

Questions?

Fragile X Syndrome & Autism

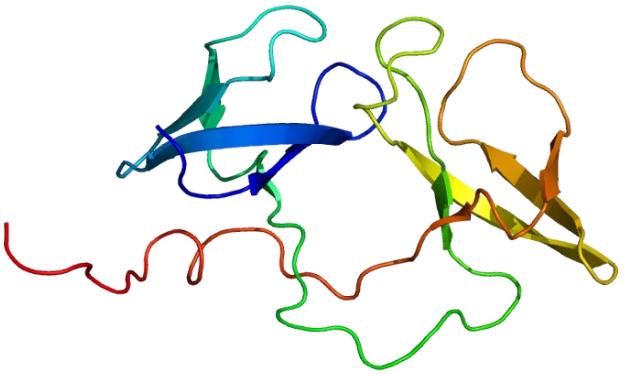
Human brain mapping 2014 – "Topological Methods Reveal High and Low Functioning Neuro-Phenotypes Within Fragile X Syndrome"

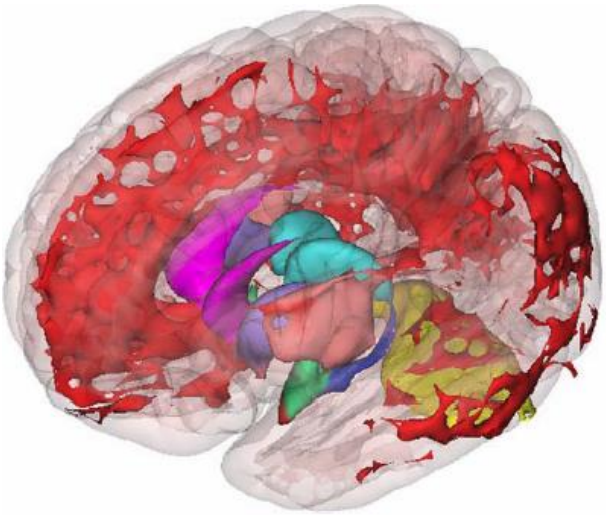
Romano, David, Monica Nicolau, Eve-Marie Quintin, Paul K. Mazaika, Amy A. Lightbody, Heather Cody Hazlett, Joseph Piven, Gunnar Carlsson, and Allan L. Reiss.

Fragile X Syndrome & Autism

- Mutations of the FMR1 gene are associated with Fragile X syndrome (FXS)
- Related to inherited cause of developmental disability and autism
- Goal: examine variation in brain structure in FXS with TDA to assess relation to IQ levels and autism-related behaviors

FMR1 gene

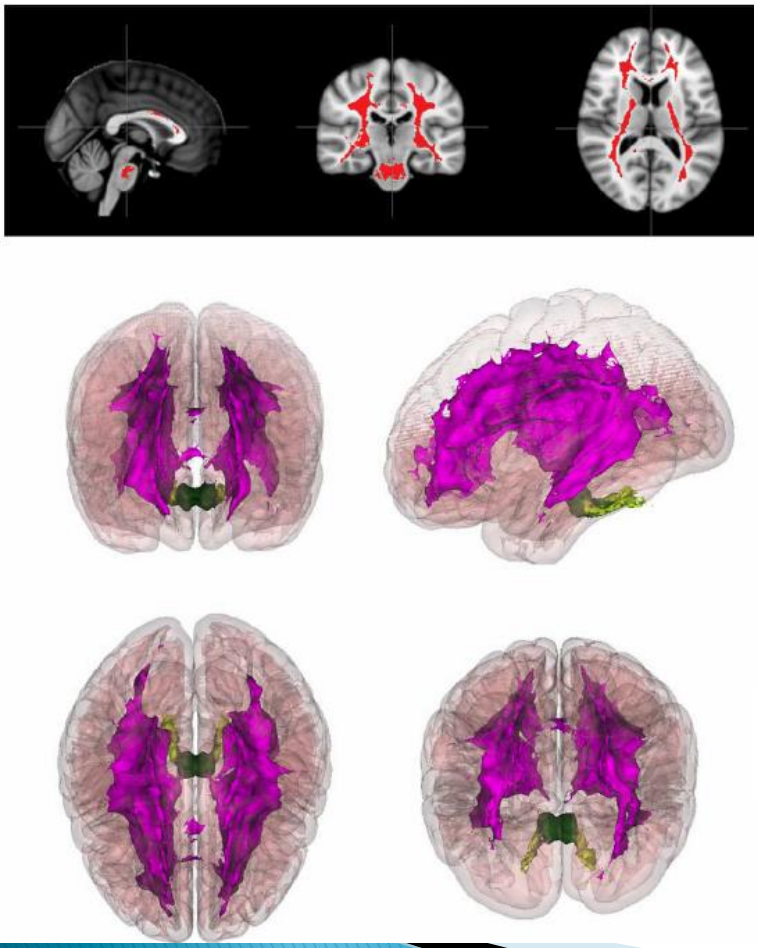




Human brain mapping 2014, Romano D., et al.

Fragile X Syndrome & Autism

- Data preprocessing:



raw MRI images

standard VBM preprocessing

normalized 1x1x1 mm images

averaging over 4x4x4 mm blocks

smoothed 4x4x4 mm images

vectorization of images

data matrix

thresholding and normalization of voxels by variance

point cloud

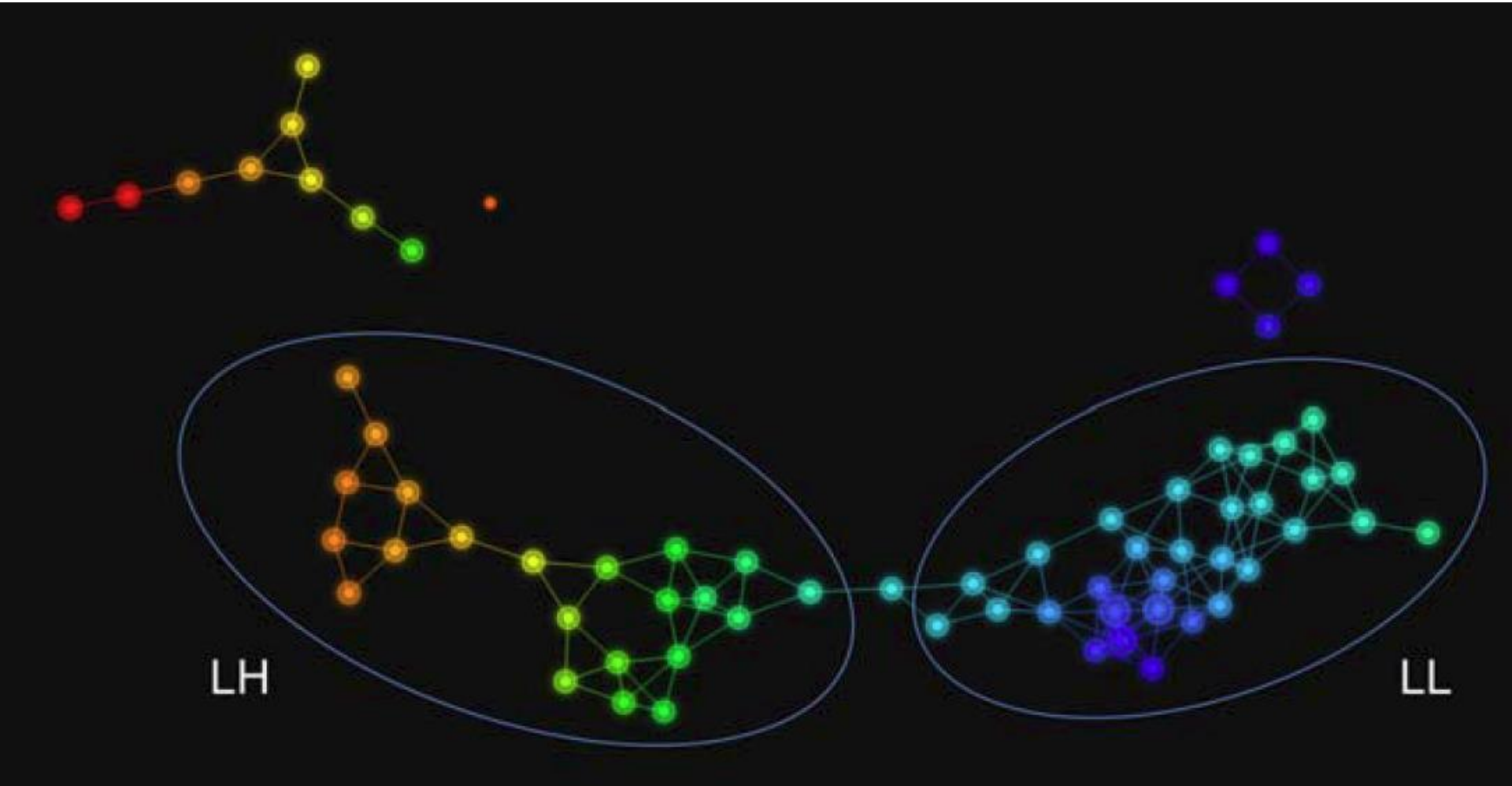
Human brain mapping 2014, Romano D., et al.

Fragile X Syndrome & Autism

- Filters: 1st and 2nd principal components

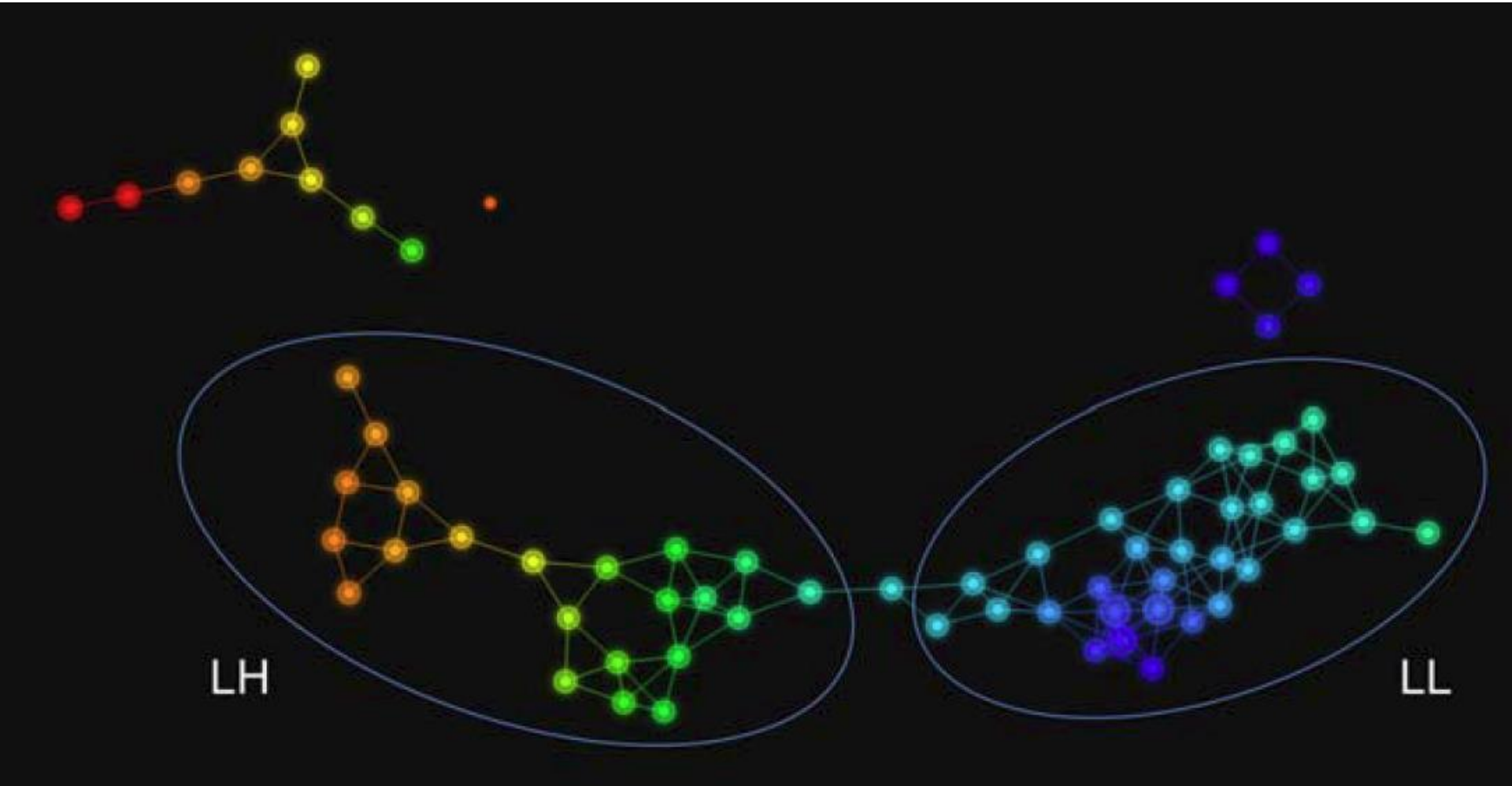
Human brain mapping 2014, Romano D., et al.

Fragile X Syndrome & Autism



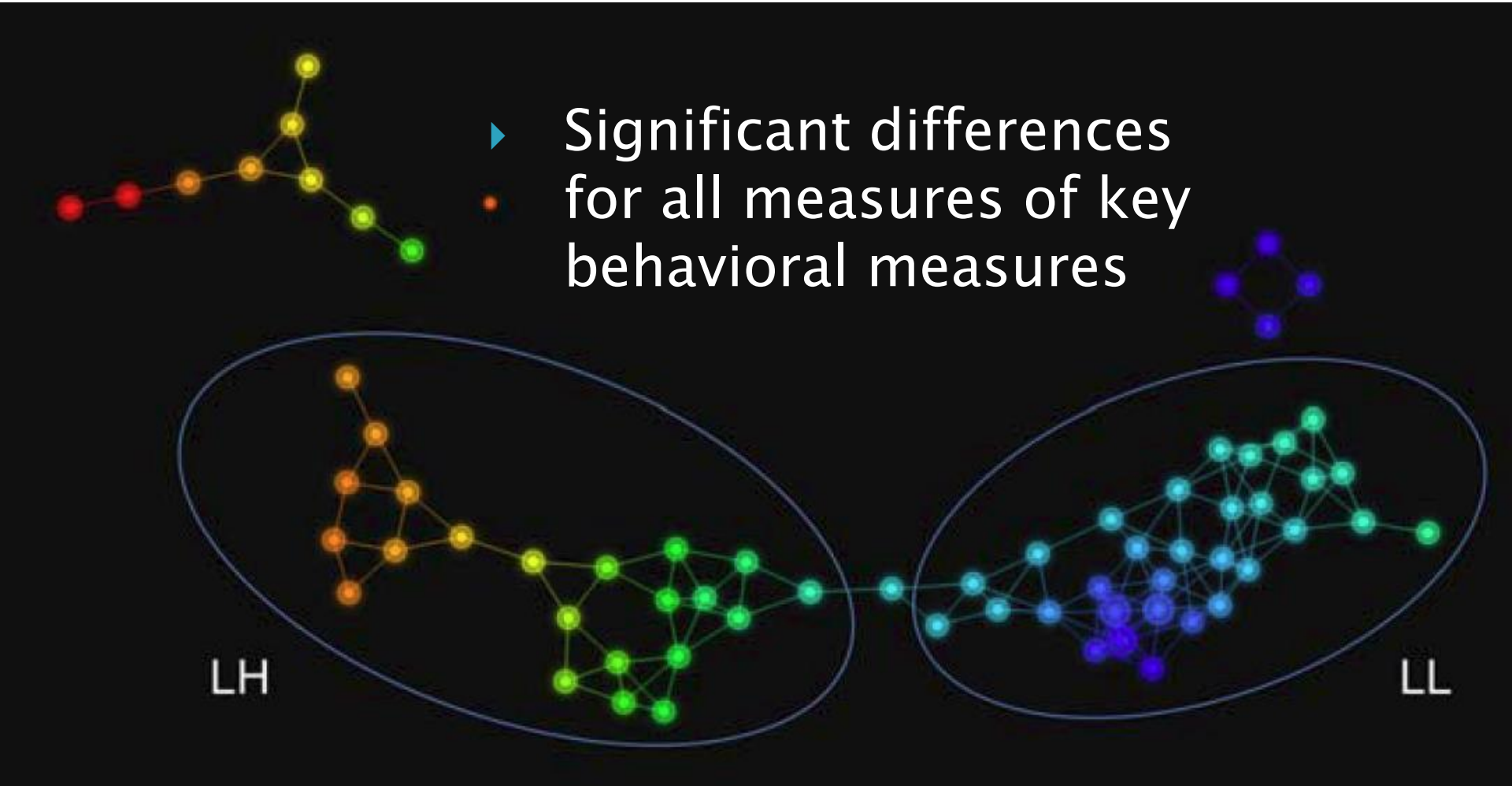
Fragile X Syndrome & Autism

- Two significantly different groups



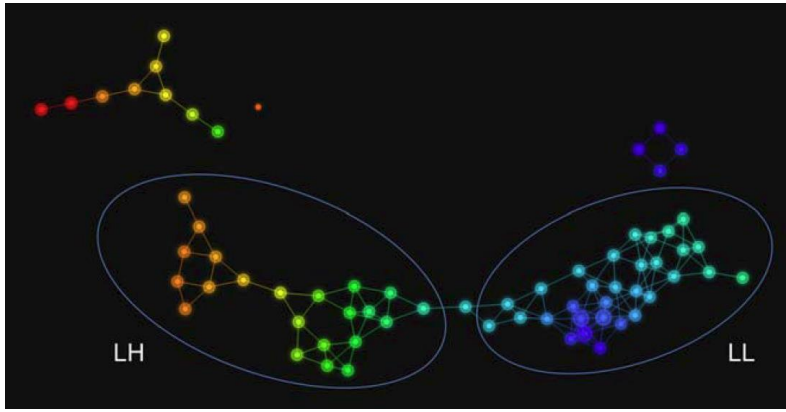
Fragile X Syndrome & Autism

- Two significantly different groups



Fragile X Syndrome & Autism

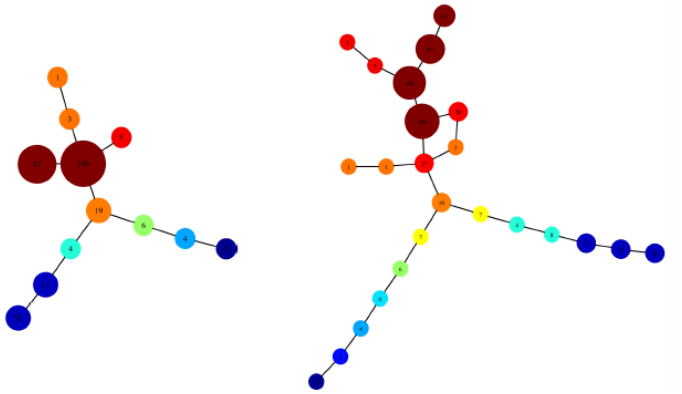
- Data preprocessing was required
- PCA components were used as filter functions
- Data visualization with TDA methodology allowed identify two significantly different FXS groups



Questions?

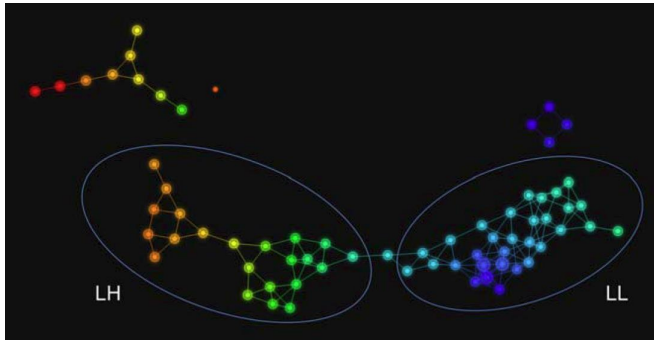
Available software

- Mapper (free)



- Code for Python
- Visualization with GraphViz

- Ayasdi's Iris (commercial)



- Includes visualization
- Has free trial period

Thank You!

Questions?

Extras

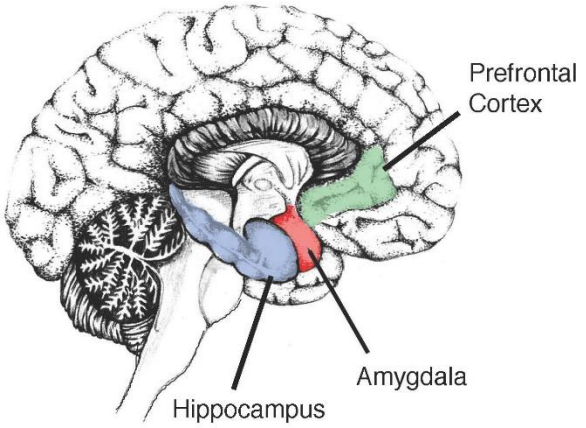
Hippocampal Map (2012)

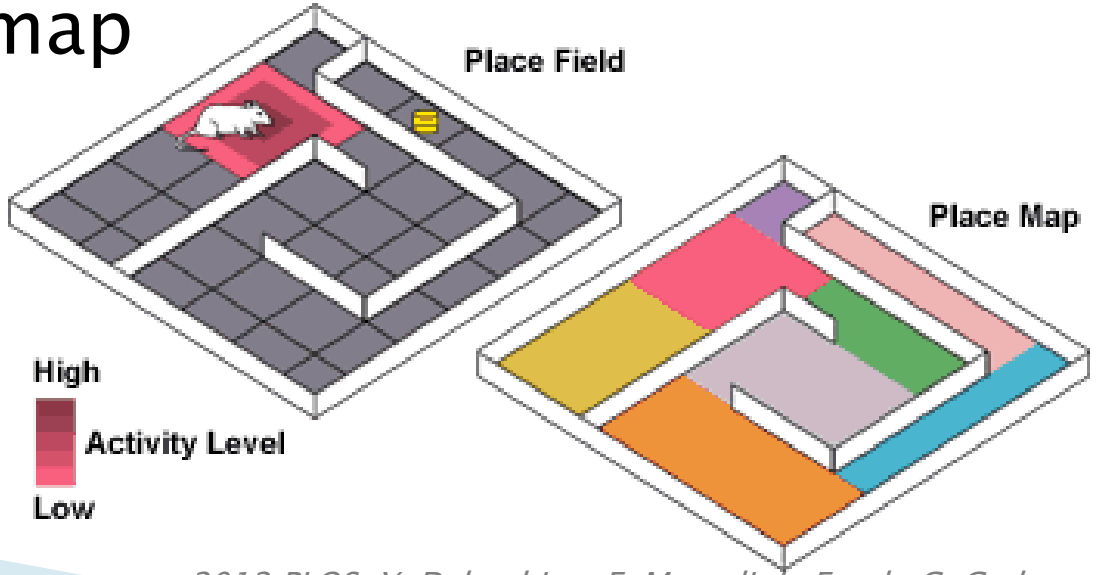
2012 PLOS – “A Topological Paradigm for Hippocampal Spatial Map Formation Using Persistent Homology”

Y. Dabaghian, F. Memoli, L. Frank, G. Carlsson

Hippocampal Map (2012)

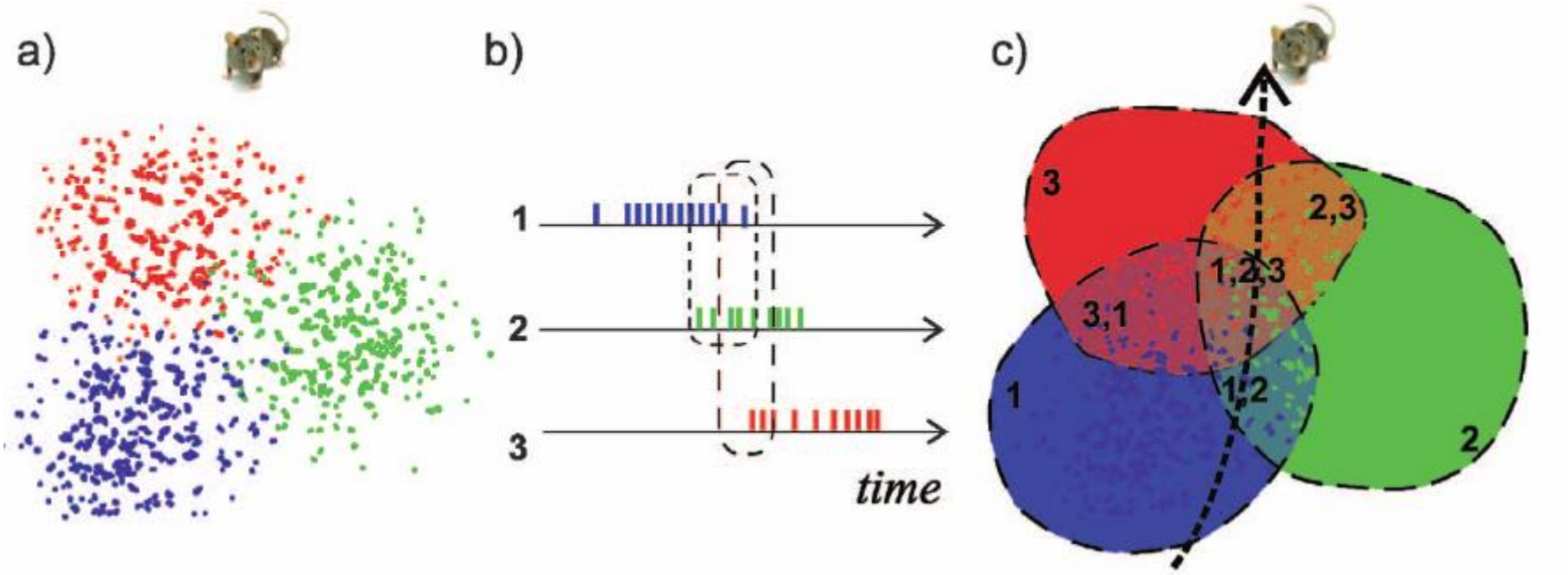
- Hippocampus plays a central role in forming internal spatial map
- Place cells get activated when an animal visits certain place
- Create an internal map
- When place cells overlap, we observe co-firing





2012 PLOS, Y. Dabaghian, F. Memoli, L. Frank, G. Carlsson

Hippocampal Map (2012)



- A map encoded by co-firing will be a topological map, i.e., based on connectivity and adjacency

2012 PLOS, Y. Dabaghian, F. Memoli, L. Frank, G. Carlsson

Hippocampal Map (2012)

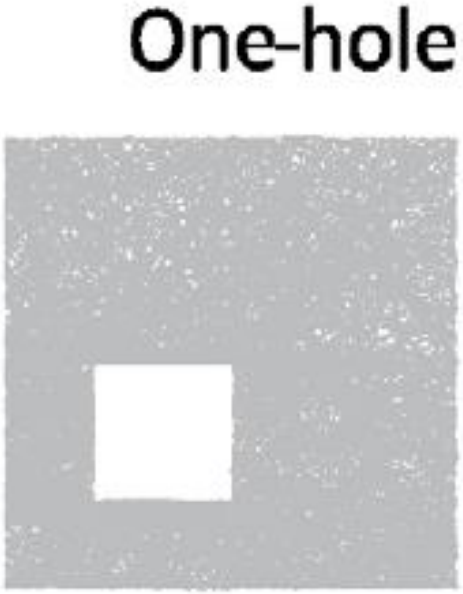
- We can reconstruct map using the basic theorem of algebraic topology:

If one covers a space X with a sufficient number of discrete regions, then it is possible to reconstruct the topology of space X from the pattern of the overlaps between the regions

- Co-firing of place cells is the key to decoding spatial information
- What is the minimal time to learn it?

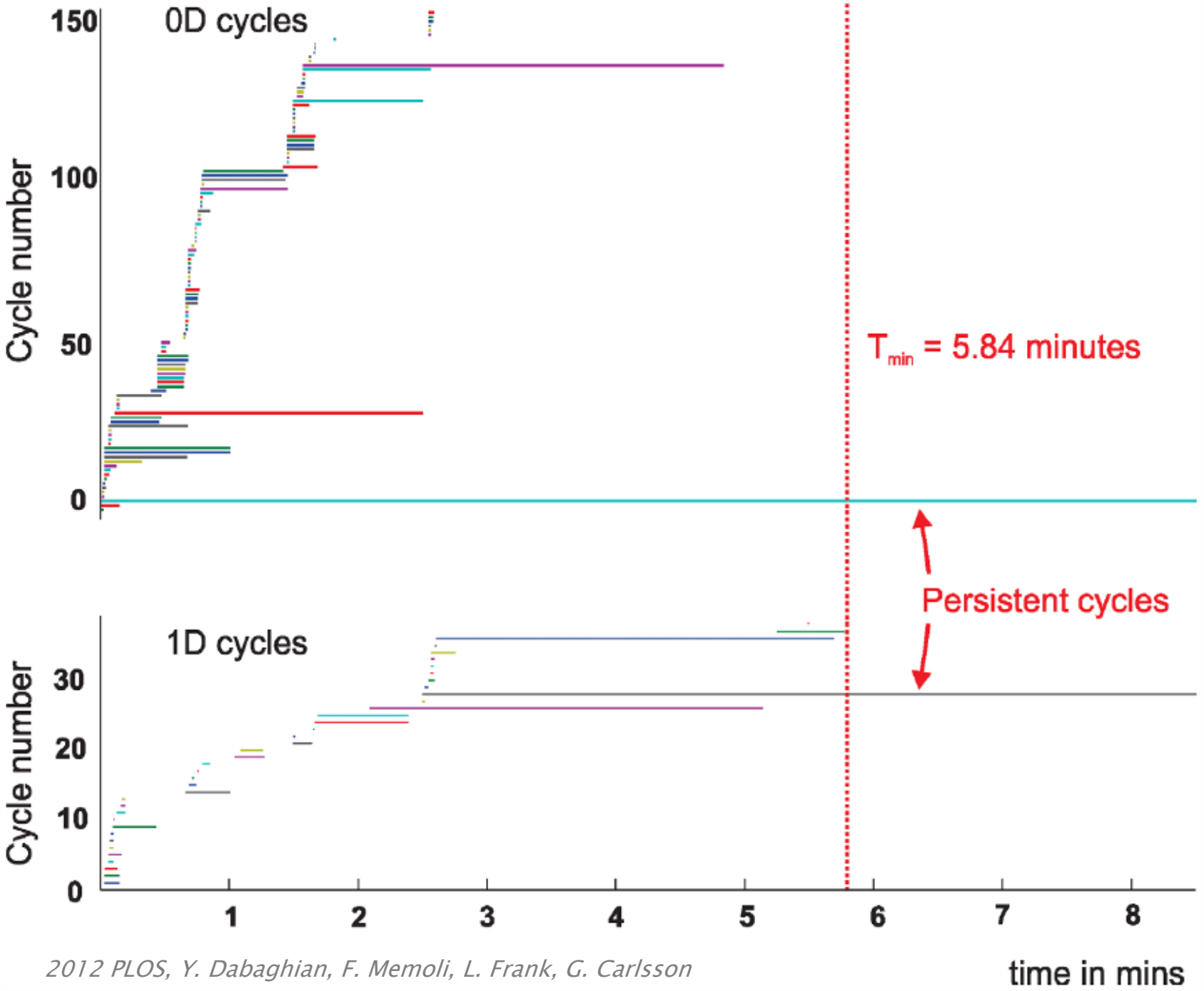
2012 PLOS, Y. Dabaghian, F. Memoli, L. Frank, G. Carlsson

Hippocampal Map (2012)



- Observe: when the space is learned, we'll have 1 connected component ($\beta_0 = 1$) and 1 loop ($\beta_1 = 1$)

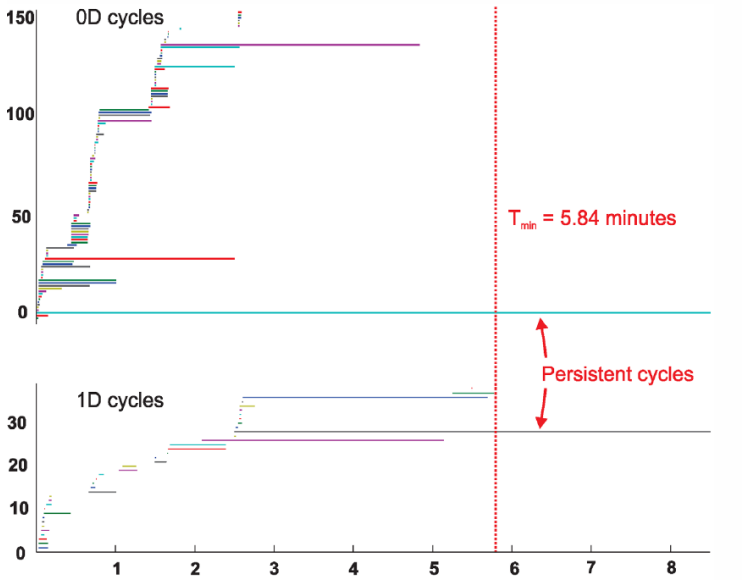
2012 PLOS, Y. Dabaghian, F. Memoli, L. Frank, G. Carlsson



2012 PLOS, Y. Dabaghian, F. Memoli, L. Frank, G. Carlsson

Hippocampal Map (2012)

- By analyzing simulated spiking showed the hippocampal place cells must operate within certain parameters of neuronal activity in order to learn the map
- Parameters vary with geometric and topological properties of the environment
- Beyond certain limit cannot form correct map



2012 PLOS, Y. Dabaghian, F. Memoli, L. Frank, G. Carlsson