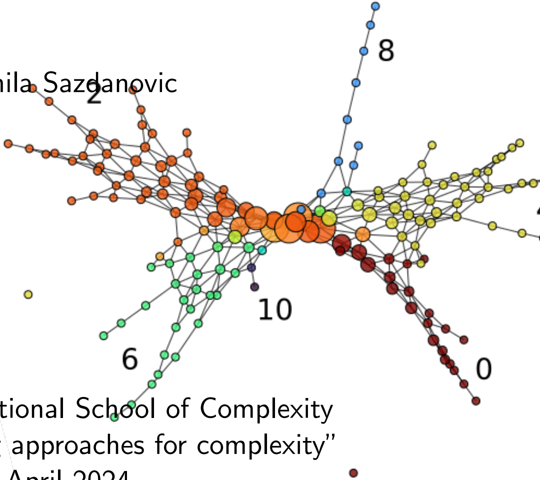
DATA, RELATIONS AND THEIR SHAPE
PART I

Radmila Sazdanovic

NC STATE UNIVERSITY



18th Erice International School of Complexity
“Machine Learning approaches for complexity”
25 April 2024

- TOPOLOGY what it is and what it can be used for?
- APPLIED ALGEBRAIC TOPOLOGY (AAT)
  - resources, software, etc.
  - interactions with other data analysis methods
  - main tools
- APPLICATIONS
  - Complex data with hidden dependencies, symmetries, etc.
    - game theory
    - materials science
    - cancer genomics
  - Theoretical mathematics: knot theory, reresentation theory

TOPOLOGY: FIELD

deformation of maps and spaces with equivalence relations such as homotopy and homeomorphism, emphasize qualitative properties.

TOPOLOGY: OF A SPACE

- A collection of all open sets in the space
- System of "neighbourhoods" that allows for notions of proximity without metric distances.
- Sufficient for defining continuity, convergence, connectivity that generalize standard notions from metric spaces.

- CHARACTERIZATION Topological features are global and qualitative; suitable for classification.
- STABILITY Topological features are robust.
- INTEGRATION Converting local data to global features
  - A graph has an Eulerian circuit iff every node has even degree
  - The Gauss-Bonnet Theorem relates the Euler characteristic to the Gaussian curvature.
- OBSTRUCTION Answering feasibility questions even when answers are hard to compute such as classes, degrees, etc.
  - Borsuk–Ulam theorem: $f : S^n \to \mathbb{R}^n$ is continuous then there exists an $x \in S^n$ such that $f(-x) = f(x)$.
  - Hairy ball theorem: there is no nonvanishing continuous tangent vector field on S^{2n}.

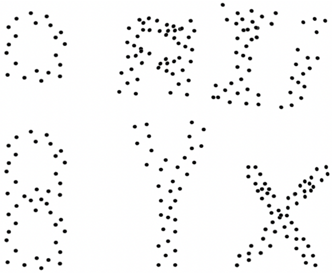
"Topology! The stratosphere of human thought! In the twenty-fourth century it might possibly be of use to someone..."

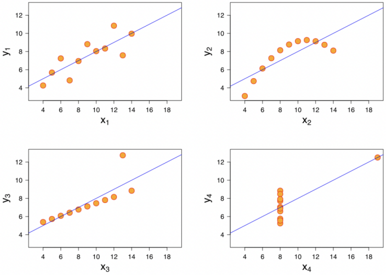
The First Circle, A. Solzhenitsyn

- Applied Algebraic Topology Network Bringing together researchers across the world to develop and use applied and computational topology: youtube, tutorials, weekly online seminars, interviews, workshops, poster sessions, etc.
- COURSE Foundations of Topological Data Analysis by R. Ghrist, V. Nanda Videos and notes
- BOOKS Computational Topology for Data Analysis T. Dey, Y. Wang; Elementary Applied Topology R. Ghrist, ...
- SOFTWARE Kepler Mapper, Ball Mapper, TDA Mapper, Dionysus, PHAT, GHUDI, Eirene, Ripser, JavaPlex,

- Software for statistical analysis of persistent homology and density clustering: R Scripts by Bubenik, R Package Fasy, Kim, Lecci, Maria, Millman, Rouvreau
- Statistical properties of topological features of data by Turner, Mileyko, Mukherjee, Harer.
- An Introduction to Topological Data Analysis: Fundamental and Practical Aspects for Data Scientists Chazal, Michel
    - consistency and the convergence rates of TDA methods.
    - confidence regions for topological features and discussing the significance of the estimated topological quantities.
    - selecting relevant scales on which the topological phenomenon should be considered, as a function of observed data.
    - dealing with outliers and providing robust methods for TDA.

1. UNDERSTANDING ML MODELS using topology (Grigsby, Lindsey and Rolnick Hidden symmetries of ReLU networks; Masden, Meyerhoff, Wu, Saul and Arendt Machine Learning Explanations with TDA, etc.)
2. TDA FEATURES AS AN INPUT TO ML Topological features can be used as an input to ML algorithms to improve accuracy e.g. Persistence Images: A Stable Vector Representation of Persistent Homology Adams
3. TOPOLOGICAL DEEP LEARNING deep learning models for data supported on topological domains
   - Graph learning models: Weisfeiler-Lehman meets Gromov-Wasserstein Chen, Lin, Memoli, Wan, Wang
   - Topological Deep Learning: Graphs, Complexes, Sheaves e.g. Message Passing Simplicial Networks Bodnar

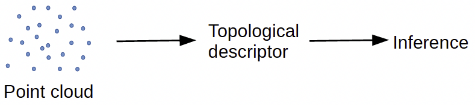




- How to design a model to fit any possible shape of the data?
- Anscombe's quartet: example of point clouds with the same descriptive statistics, but very different shapes.
- DATA HAS SHAPE, SHAPE CARRIES INFORMATION

WHY TOPOLOGY

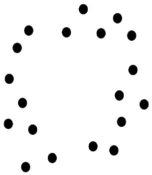
- Data comes in different forms: graphs, grids, manifolds
- Unifying properties
  - localized
  - relational: some notion of proximity



TOPOLOGICAL DATA ANALYSIS TOOLS

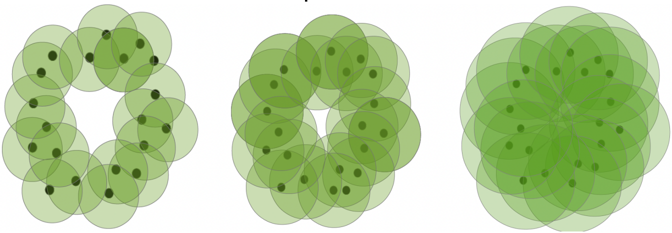
- PERSISTENCE integer, bar code, function/curve, multi-dimensional persistence.
- MAPPER ALGORITHMS
  - Mapper - Singh, Mémoli and Carlsson (2007)
  - BallMapper - Dłotko (2019)

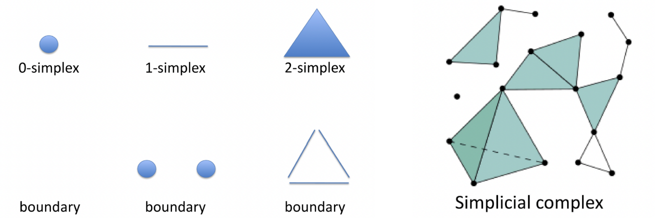
HOW TO EXTRACT TOPOLOGY FROM A POINT CLOUD?



- What do you see?
- We may say that we see a circle, but we really see 19 points ... that may be sampled from a probability distribution supported at a circle.

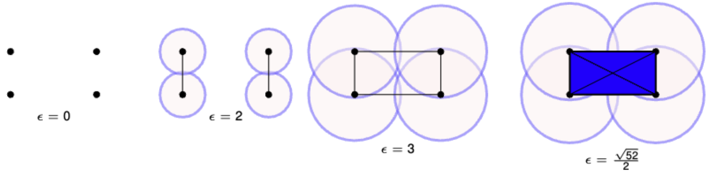
First: construct a cover by choosing a range of ϵ_is and a ball of that radius around each data point.





- n-simplex σ_n
- Simplicial complex: combinatorial representation of topological spaces built out of simple pieces: points, edges, triangles, tetrahedra, etc.
- Higher dimensional analogue of graphs
- Boundary of simplex $\partial_n(\sigma_n)$ consists of all (n-1)-simplices (faces) in its topological boundary

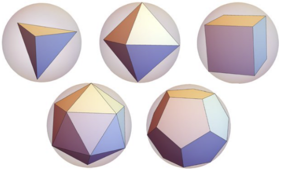
EULER CHARACTERISTICS χ



$\chi = 4$ $\quad\chi = 4 - 2 = 2$ $\quad\chi = 4 - 4 = 0$ $\quad\chi = 4 - 6 + 4 = 2$

$$\chi = \#points - \#edges + \#faces - \ldots = \sum_{i \geq 0} \#(\mathrm{i} - \mathrm{dim\,cells})$$

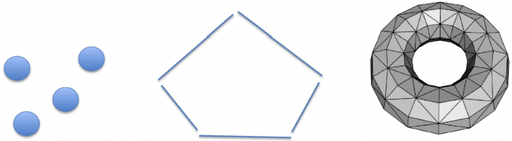
- χ is a topological invariant
- Given a graph G, $\chi(G) = 1$ if and only if G is a tree.



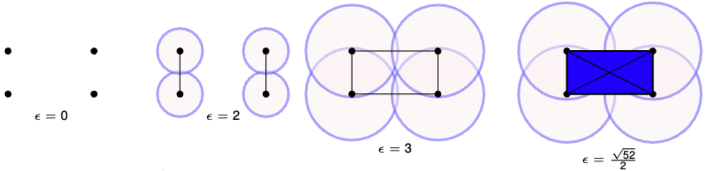
- $\chi(\mathrm{tetrahedron}) = 4 - 6 + 4 = 2$
- $\chi(\mathrm{cube}) = 8 - 12 + 6 = 2$
- $\chi(\mathrm{convex\,polyhedron}) = 2$ because they are all topologically equivalent to the 2–dim ball S^2

$X \to$ CHAIN COMPLEX $C_\star(X) \to$ HOMOLOGY $H_\star(X)$

- Given a simplicial complex X an **n-chain**, $c = \sum_i a_i \sigma_i$ is a formal sum of n-simplices σ_i in X with $a_i \in \mathbb{Z}$.
- CHAIN COMPLEX $C_\star(X)$ consists of n–th CHAIN GROUP $C_n(X)$ free, abelian, generated by all n-simplices and the BOUNDARY MAP $\partial_n : C_n(X) \to C_{n-1}(X)$ such that $\partial_n \partial_{n+1} = 0$
- n–th HOMOLOGY GROUP $H_n(X) = \ker(\partial_n)/\operatorname{im}(\partial_{n+1})$.
- n–th BETTI NUMBER $\beta_n(X) = \operatorname{rk}(H_n(X))$.

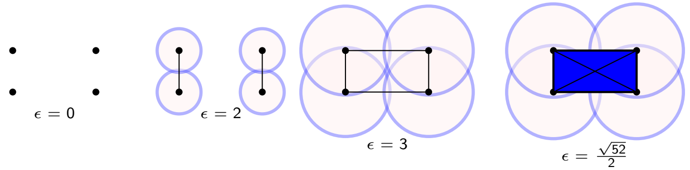


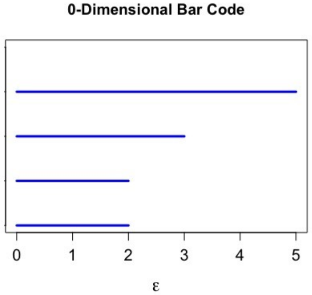
0-cycle 1-cycle 2-cycle

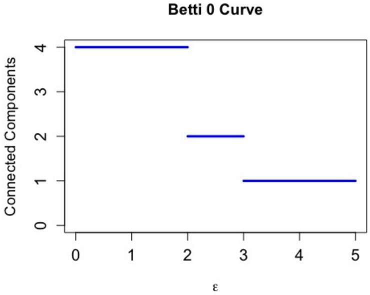


- $X \subseteq \mathbb{R}^n$ a finite point cloud, $\epsilon > 0$
- VIETORIS-RIPS COMPLEX $VR(X, \epsilon)$ simplicial complex with a face $\{x_1, \ldots, x_k\} \in X$ if and only if $B(x_i, \epsilon) \cap B(x_j, \epsilon) \neq \emptyset$ for all $i, j \in [k]$.
- FILTRATION OF X nested sequence $X = X_0 \subseteq X_1 \subseteq \ldots \subseteq X_k$ of simplicial complexes $X_i = VR(X, \epsilon_i)$ for $\epsilon_0 = 0$, $\epsilon_i < \epsilon_{i+1}$
- PERSISTENCE OF X: $H_\star(X_0) \to H_\star(X_1) \to \ldots \to H_\star(X_k)$

BAR CODES AND BETTI CURVES (DIMENSION 0)

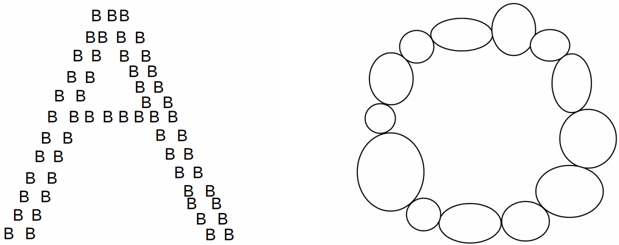






- Capture persistence of topological features through filtration
- In dimension 0: tracking the number of connected components

PERSISTENT HOMOLOGY



- Robust, stable, multi scale, coordinate–free, compressed, tool to detect connected components, cycles, voids etc.
- Output: bar code (persistence diagram), Betti curve, persistence images, landscapes - all can be compared and used to distinguish data sets

- Cancer is a polygenic disease: genomic events are selected in order to produce a sophisticated and coordinated outcome
- Data: Horlings and TCGA
- Luminal A: ER and or PR+; Low grade, slow growing; Best prognosis, responds to hormone therapy
- Luminal B: ER+, can be PR-; Intermediate/high grade, grows faster than Lum A; Worse prognosis than Luminal A, responds to hormone and chemo therapy
- HER2+: ER/PR-; More aggressive than luminals; Responds to chemo and some HER2 specific therapies
- Basal: ER, PR, HER2-, aggressive, responds to chemo

- "Applications of topological data analysis in oncology" by Bukkuri, Andor, Darcy
- "Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival", Nicolau, Levine, Carlsson
- "Identification of relevant genetic alterations in cancer using topological data analysis", Rabadán at al.
- A series of 5 papers that apply persistent homology to cancer genomics Dewoskin at al., Arsuaga at al. 2012, Arsuaga at al. 2015, Ardanza 2016 at al. , Gonzalez 2020

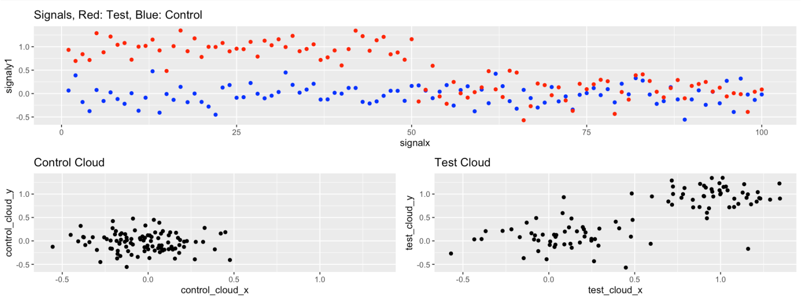
DETECTING ONCOGENES BASED ON CNA

- Copy Number Aberrations (CNAs) and gene expression relate to the breast cancer types and prognosis
- GOAL Detect co-occurring events
- COLLABORATORS:J. Arsuaga, S. Ardanza-Trevijano, J. Aslam, G. Gonzalez, A. Ushakova, J. Xiong

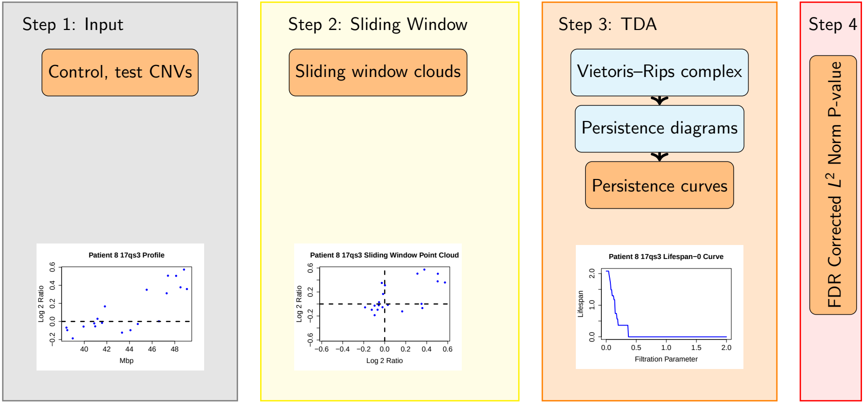
	LOW EXPRESSION	HIGH EXPRESSION
LOW COPY NUMBER	Tumor Suppressor, Cancer-Related Gene	Not Driven By Copy-Number
HIGH COPY NUMBER	Not Driven By Copy-Number	Oncogene Cancer-Related Gene

SLIDING WINDOW POINT CLOUD

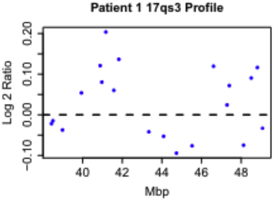
- $X = \{x_1, \ldots x_n\}$ be a time series
- Sliding window point cloud of X with window size 2 is $(x_1, x_2), (x_2, x_3) \ldots (x_{n-1}, x_n),\ (x_n, x_1)$.



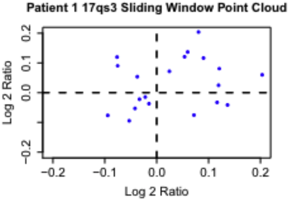
GENETIC ASSOCIATION STUDY: TAaCGH BY ARSUAGA AT AL.



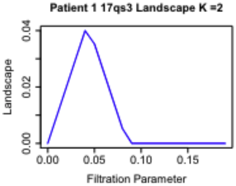
- Statistical methods on topological measurements
- TEST: breast cancer subtype
- CONTROL: the other breast cancer subtypes


Patient 1 17qs3 Profile
Log 2 Ratio
−0.10
0.00
0.10
0.20
40
42
44
46
48
Mbp

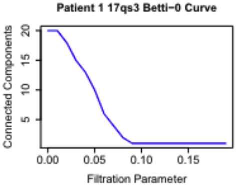
(a)


Patient 1 17qs3 Sliding Window Point Cloud
Log 2 Ratio
−0.2
0.0
0.1
0.2
−0.2
−0.1
0.0
0.1
0.2
Log 2 Ratio

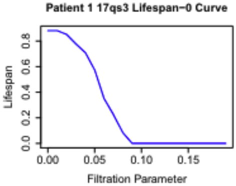
(b)


Patient 1 17qs3 Landscape K =2
Landscape
0.00
0.02
0.04
0.00
0.05
0.10
0.15
Filtration Parameter

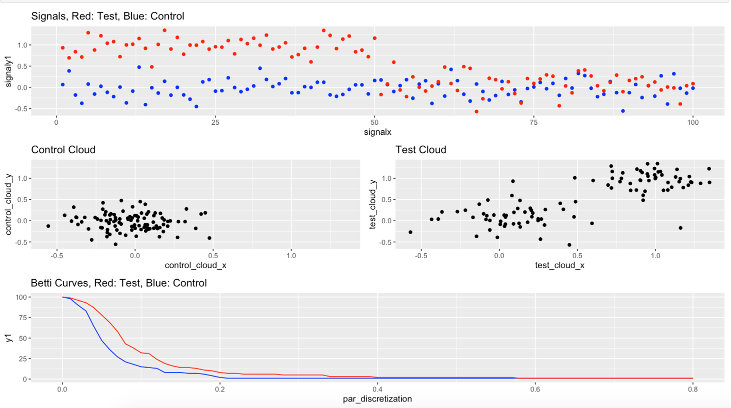
(c)


Patient 1 17qs3 Betti−0 Curve
Connected Components
5
10
15
20
0.00
0.05
0.10
0.15
Filtration Parameter

(d)


Patient 1 17qs3 Lifespan−0 Curve
Lifespan
0.0
0.2
0.4
0.6
0.8
0.00
0.05
0.10
0.15
Filtration Parameter

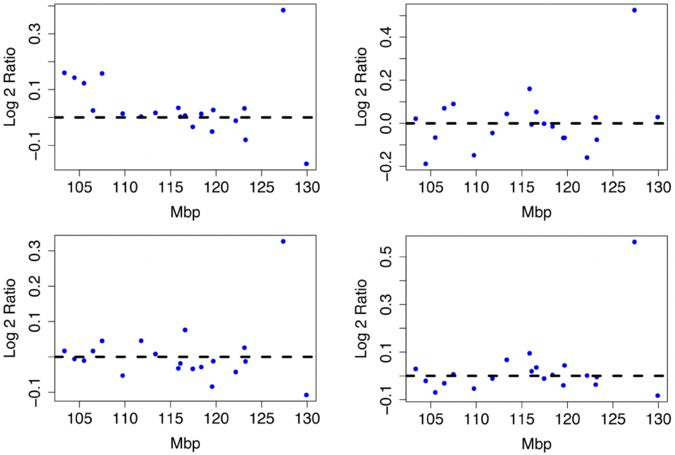
(e)



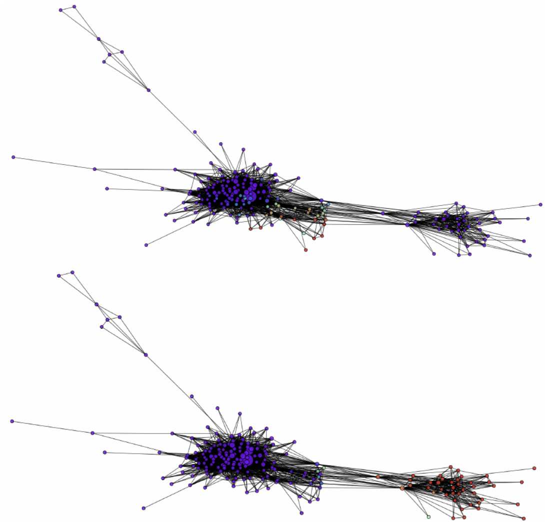
- Luminal A is characterized by gains of chromosome 1q, loss of 16p, and less common CNA changes in 8p, 8q, 11q, and 13q
- TAaCGH found 2q12.1-2q21.1 and 5p14.3-p12 as two new significant regions associated to the Luminal A subtype for further statistical analysis related to prognosis and treatment.

BACK TO DATA

Luminal A Horlings patients have a gain at 125-135 Mbp.

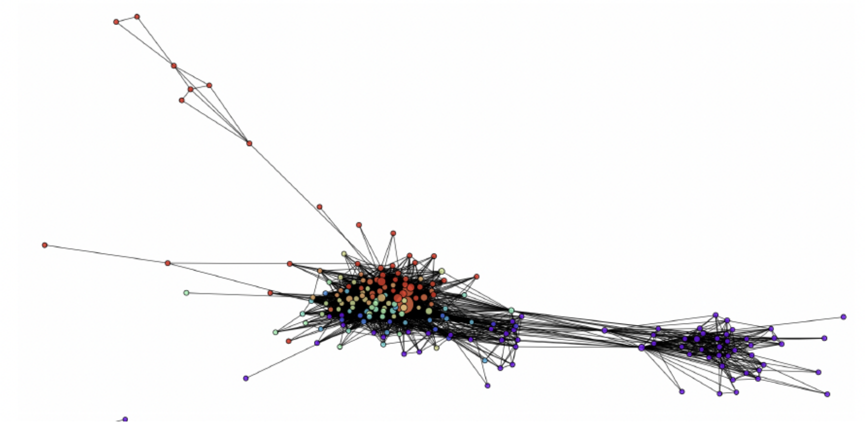


Identified segments are then used as predictor variables to build machine learning models which classify patients as one of the four subtypes



Her2

Basal



Luminal A

- Euler characteristic: enumerative topological invariant
- Homology: algebraic compression of data to its most essential features. A categorification of the Euler characteristic.
- Homology is functorial: allows tracking the way data changes, and maps between the spaces via the induced algebra maps.
- Betti numbers, Betti and Euler curves, landscapes, persistence images: topological summaries derived from homology
- DONUT Database of Original & Non-Theoretical Uses of Topology (papers, software)
- Applications include: material sciences e.g. Quantifying similarity of pore-geometry in nanoporous materials, cosmology e.g The topology of the cosmic web in terms of persistent Betti numbers, and modelling network dynamics Opinion dynamics on discourse sheaves.

QUESTIONS?