Data, their shape, and what we can learn from it

Paweł Dłotko

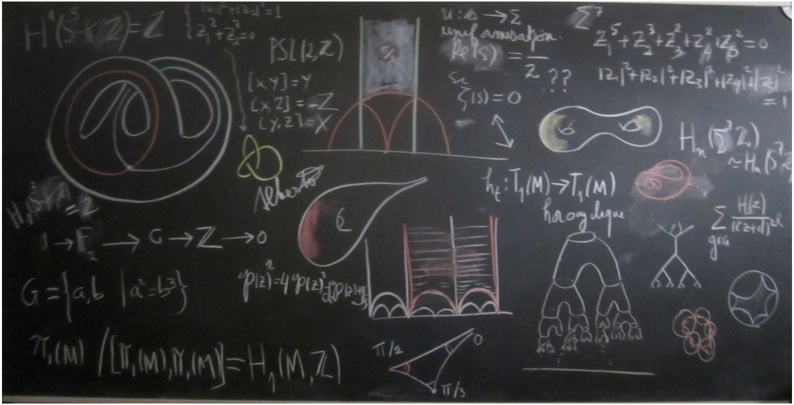
Dioscuri Centre for Topological Data Analysis, Mathematics, PAS

Linkoping 21 Nov. 2023

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

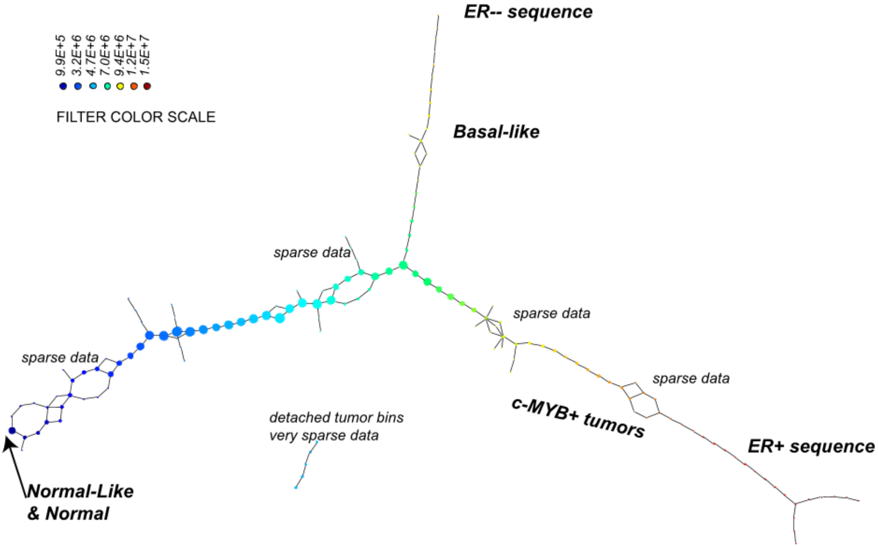
Aleksander Sołżenicyn - In the First Circle

Classical algebraic topology



By Mathieu Rémy and Sylvain Lumbroso

Applied algebraic topology



Gunnar Carlson at al. PNAS

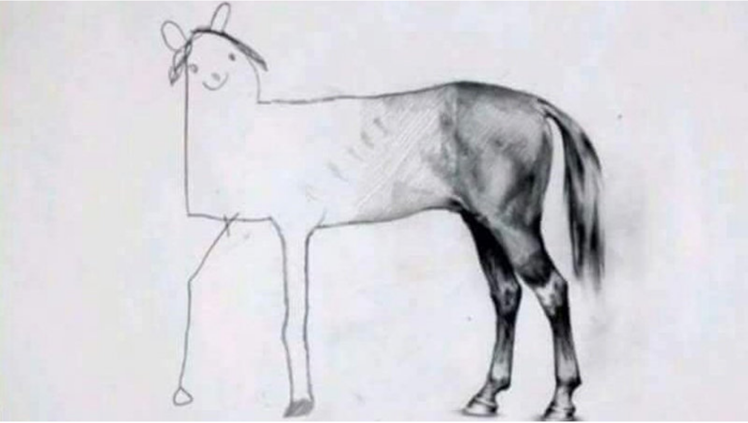
The essence of topology

Invariance to continuous deformations
Mug and torus, Wikipedia

Invariance to continuous deformations, takeaway

Topologist cannot tell apart torus from a coffee mug

Invaraince to continuous deformations = robustness to noise

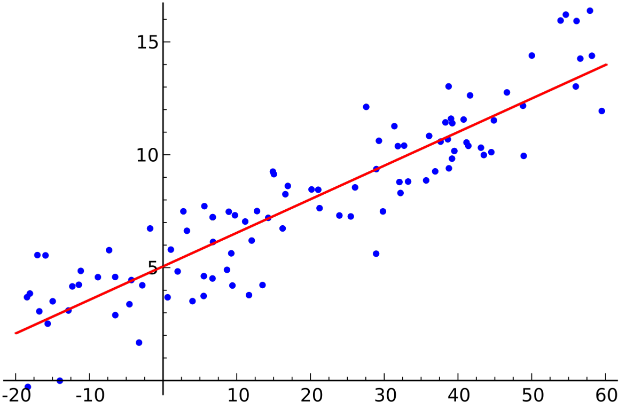


Ali Bati, unfinished horse

The credo

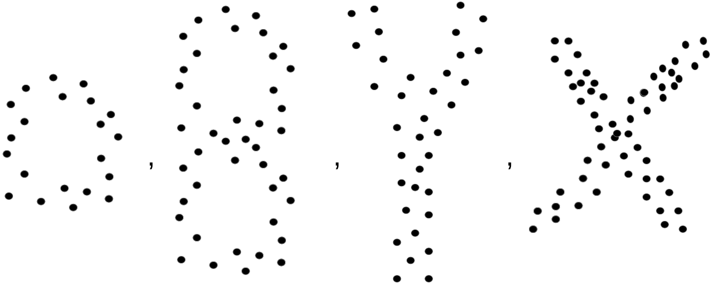
Data have shape,
shape has meaning,
meaning brings value.

We all know this story



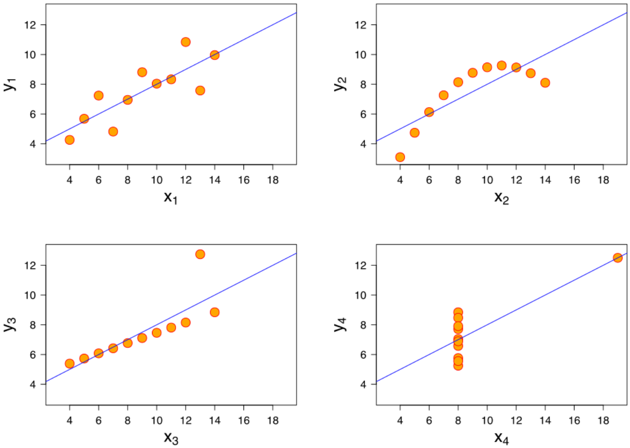
Data of a shape of a line (segment) $\rightarrow$ linear regression works

Zoology of shapes



What is the shape of our data?
How not to overfit?

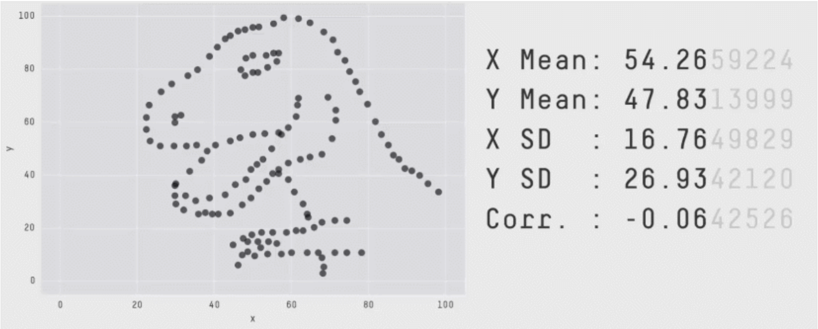
Summary statistics do not suffice, always visualize!



Anscombe's Quartet; Same statistics, different shapes

Anscombe, "Graphs in Statistical Analysis", American Statistician, 1973.

Datasaurus Dataset



Same statistics, different shapes
Alberto Cairo, `https://itsalocke.com/datasaurus/`

Topology and statistics, together

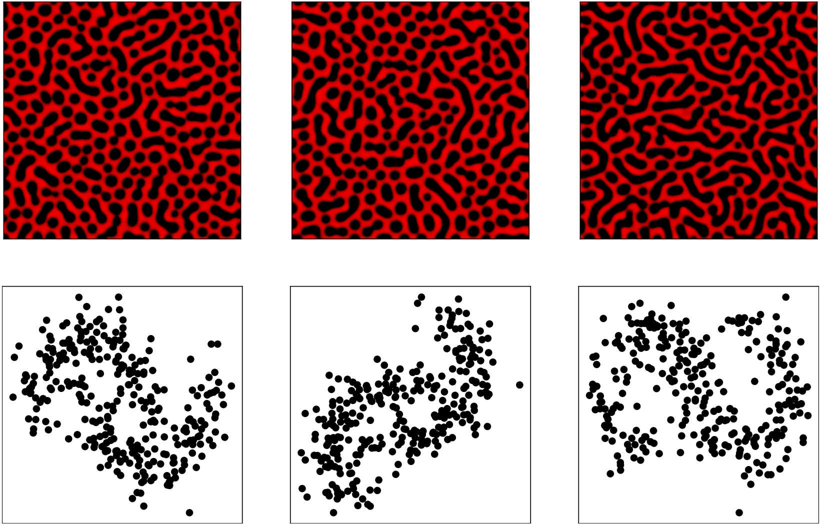
- Visualizing data brings a new level of undentstanding,
- Descriptors of shapes open up standard statistics and Machine Learning to new types of inputs.
- What if it is high dimensional, complex, not a point cloud?
- Topological invariants come to the rescue!

Quick schedule for today

- Persistent homology
- Mapper (visualization)
- ECC (descriptors)
- Topotests (blend of two dyscyplines)
- Some applications

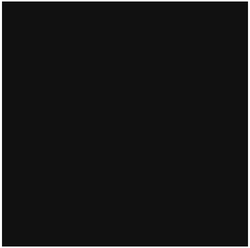
Persistent homology and learning

Quantification of a shape



Spinodal decomposition in alloys

50/50

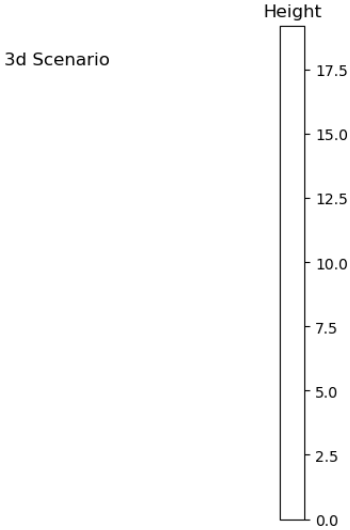


60/40

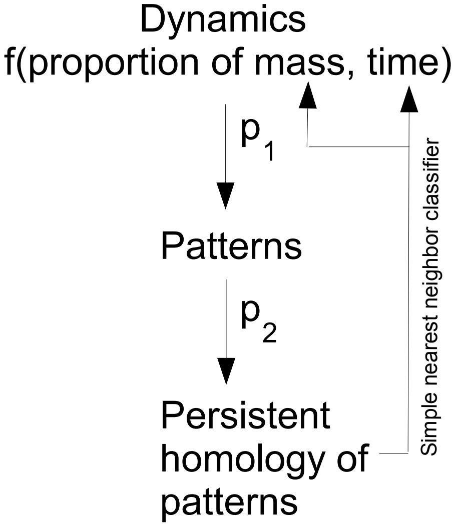
75/25

(Joint with Thomas Wanner)

Persistent homology (sublevel sets of function)

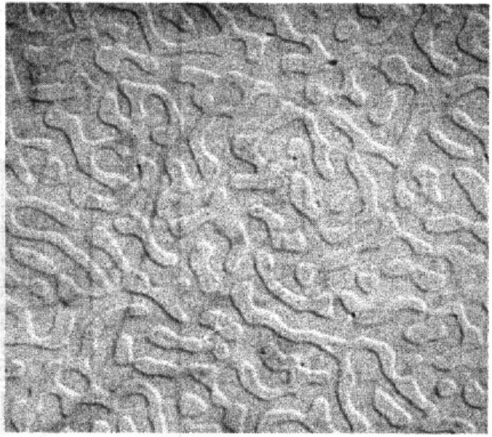


So what?

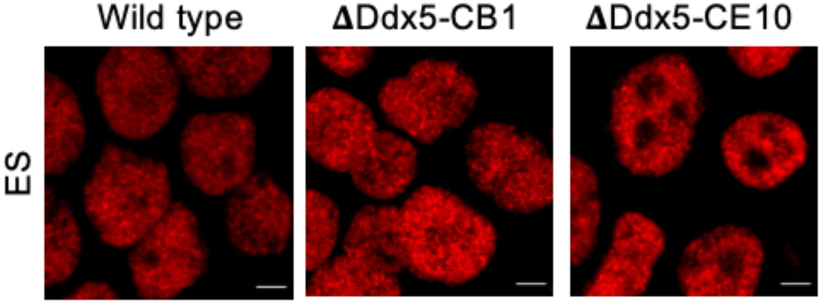

Dynamics
f(proportion of mass, time)
p1
Patterns
p2
Persistent homology of patterns
Simple nearest neighbor classifier

How can we use it in practice?

- Comparison of different models
- Comparison to the to real data.



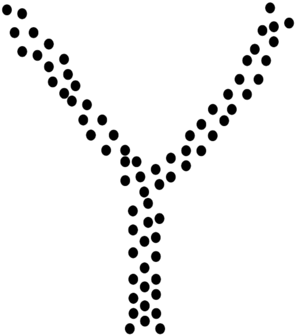
Phase separation everywhere



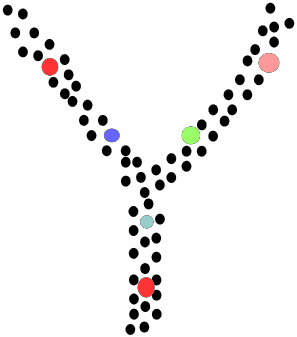
CTFC in cells

Ball mapper

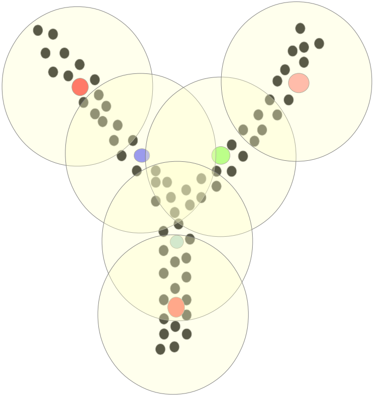
Ball Mapper algorithm



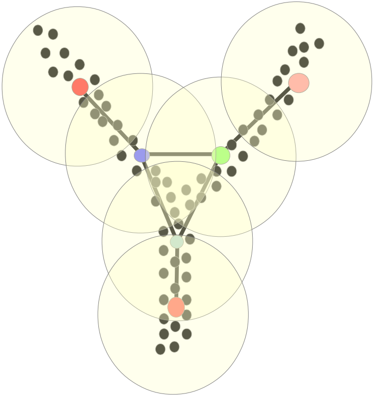
Ball Mapper algorithm



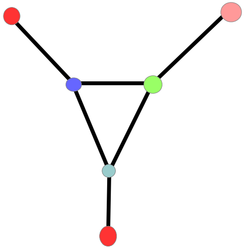
Ball Mapper algorithm



Ball Mapper algorithm



Ball Mapper algorithm

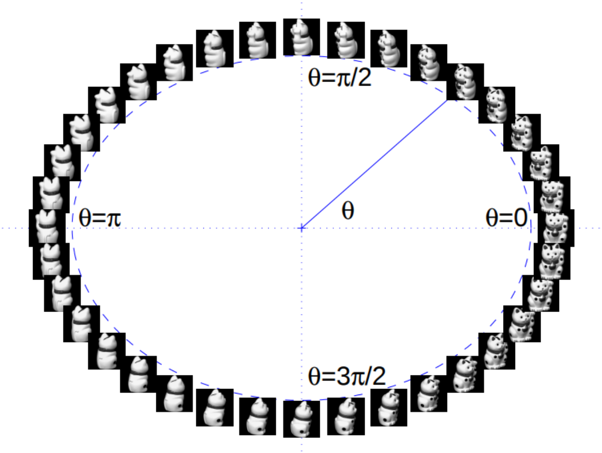


Preservation of local neighborhood, shape up to continuous deformation

Network based landscapes of data

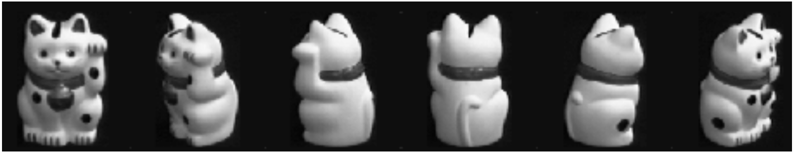
Meet the Lucky Cat

Network based landscapes of data



$128 \times 128 = 16384$ dimensional space

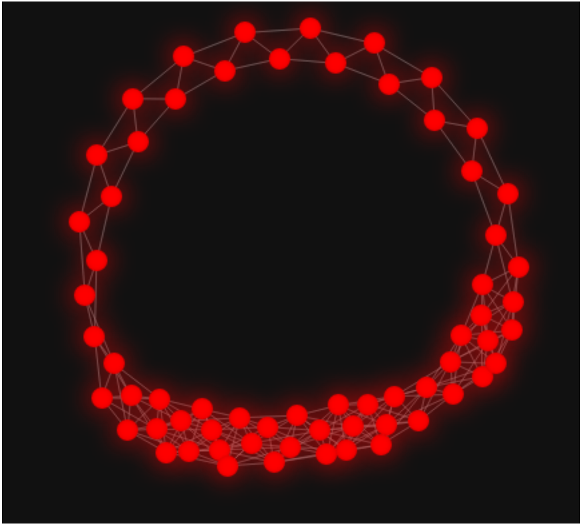
From a gray scale image to a point



$(p^1,\ldots,p^1_{16384})$ $(p^2,\ldots,p^2_{16384})$ $(p^3,\ldots,p^3_{16384})$ $(p^4,\ldots,p^4_{16384})$ $(p^5,\ldots,p^5_{16384})$ $(p^6,\ldots,p^6_{16384})$

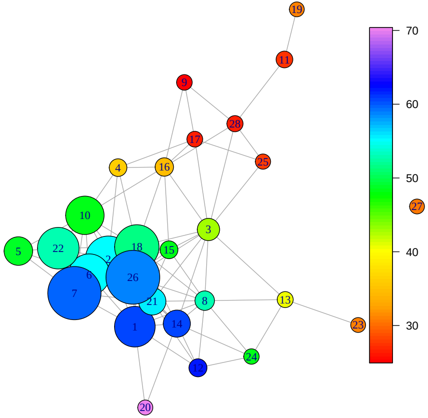
Gray scale images converted to vectors in high dimensional space

Network based landscapes of data

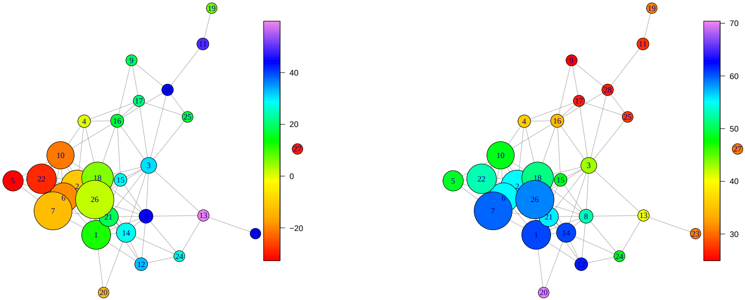


$128 \times 128 = 16384$ dimensional space

Support for Brexit in the 2016 referendum

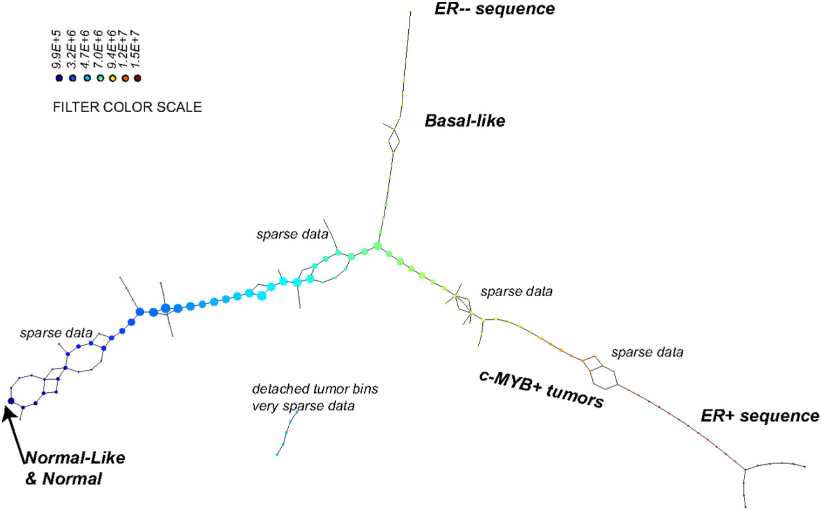


Labour vs Brexit



This is why we do not see Jeremy Corbyn anymore...

NKI, Carlson and coauthors



High dimensional noisy data

Ambient space

(~2000 dimension)

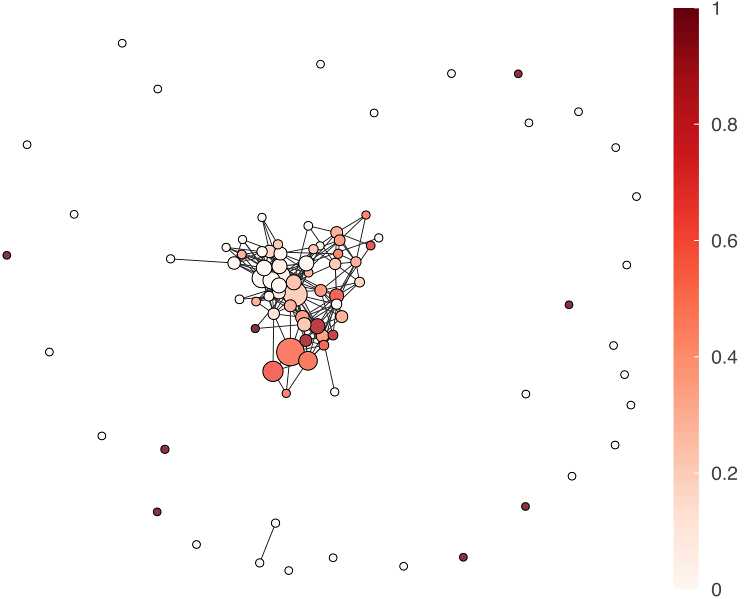
Diffusion embedding

Understanding structure of projection

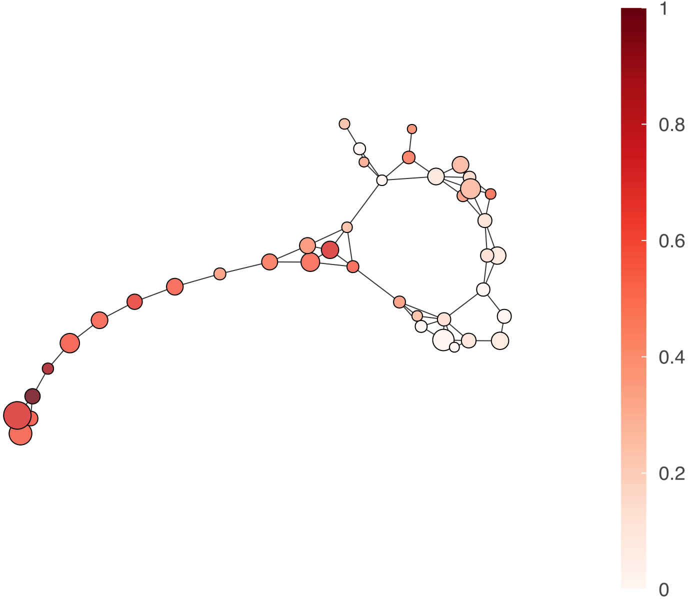
Lower dimensional representation

(~400 dimension)

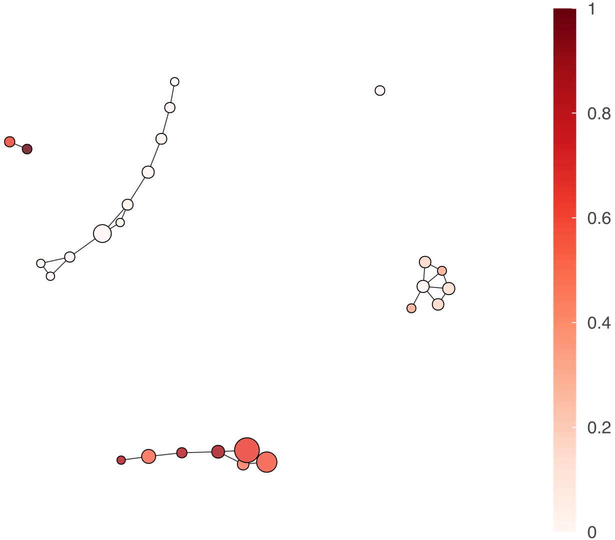
NKI, ambient dimension, BM



NKI, umam projection, BM



NKI, MoBM



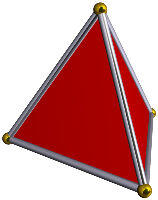
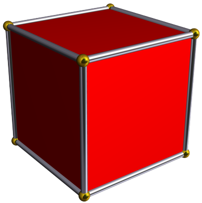
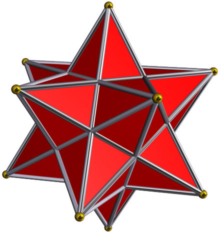
Initial collaboration with National Cancer Institute

Euler curves (and profiles)

How to encapsulate information about shape?

- Classical - homology, persistent homology,
- New - Euler characteristic curves and profiles,
- New - Characteristics of merging structure of points,...

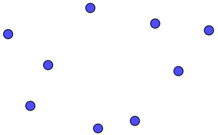
Answer: the Euler Characteristic!

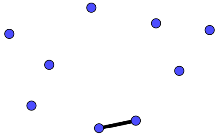
$$\chi(P) = V - E + F$$

Image source: Wikipedia

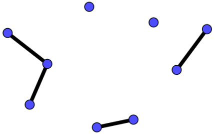
Euler characteristics, point clouds



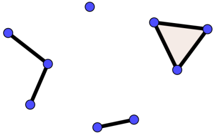
(a) $\chi = 9$



(b) $\chi = 9 - 1 = 8$

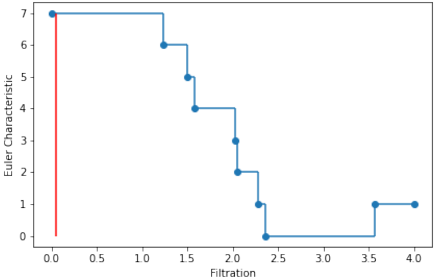


(c) $\chi = 9 - 4 = 5$



(d) $\chi = 9 - 6 + 1 = 4$

Euler Characteristic Curve - Example

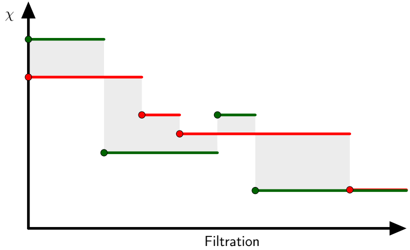


Distance between ECCs

Definition

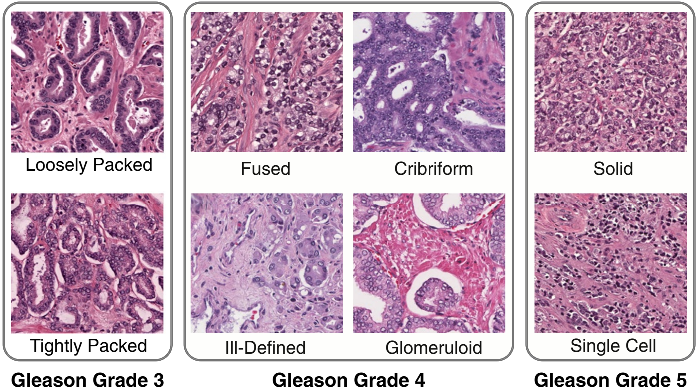
Let K_1 and K_2 be two filtered cell complexes. The L_1 *distance between* their Euler Characteristic Curves is

$$||ECC(K_1,t) - ECC(K_2,t)||_1 = \int_{\mathbb{R}} |ECC(K_1,t) - ECC(K_2,t)|dt \quad .$$

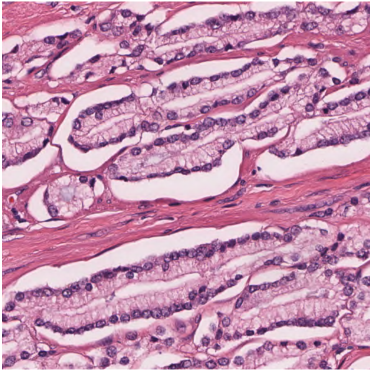


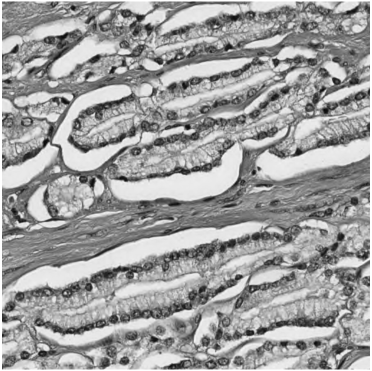
Two Euler Characteristic Curves in red and green. The absolute value of their difference is highlighted in shaded gray.

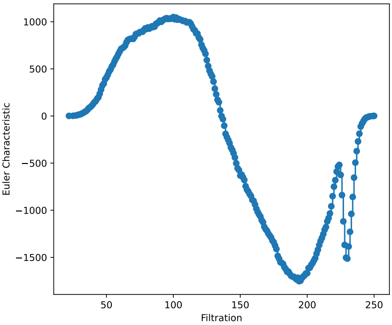
Medical applications - Histology



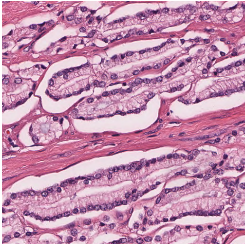
Lawson, Sholl, Brown et al. Persistent Homology for the Quantitative Evaluation of Architectural Features in Prostate Cancer Histology. 2019



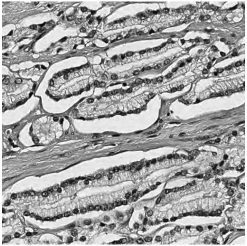




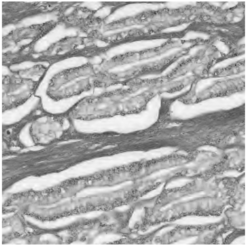
Euler Characteristic
Filtration
1000
500
0
−500
−1000
−1500
50
100
150
200
250



Full image

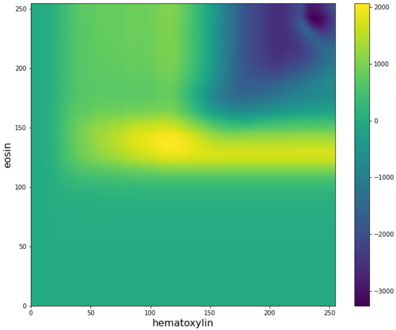


Hematoxylin

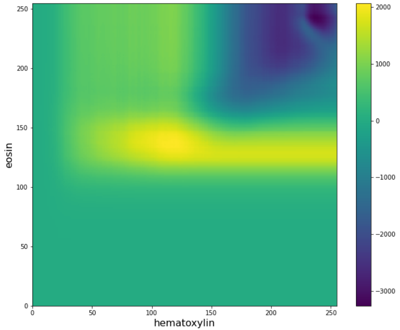


Eosin

Euler Characteristic Profiles



Euler Characteristic Profiles



hematoxylin ECC	hematoxylin & eosin ECP
0.765 ± 0.001	0.826 ± 0.001

Mean test accuracy for the Gleason 3 vs Gleason 4 classification using ECCs or ECPs as input to an SVM classifier.

Topological goodness of fit tests (topotests)

Introduction

One- and two-sample tests

- **One-sample problem**: We are given a data sample $X = \{x_1, x_2, \ldots, x_n\}, x_i \in R^d$ and cumulative distribution function $F : R^d \to [0, 1]$. Does the data X follow the distribution F: $X \sim F$?

$$H_0 : X \sim F \quad \text{vs.} \quad H_1 : X \nsim F$$

- **Two-sample problem**: We are given two samples $X_1 \sim F_1$ and $X_2 \sim F_2$ and want to test hypothesis that X_1 and X_2 were drawn from the same (unknown) distribution

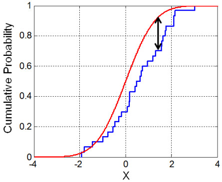
$$H_0 : F_1 = F_2 \quad \text{vs.} \quad H_1 : F_1 \neq F_2$$

Testing

Available methods depend on the data dimension (for one-sample problem)

- 1-D: plenty of available tests: e.g. Kolmogorov-Smirnov, Cramer-von Mises, Anderson–Darling, Chi-squared, Shapiro-Wilks
- 2-D: theoretical results for Kolmogorov-Smirnov and Cramer-von Mises, some implementations available in `python` and `R`
- *d*-D: Kolmogorov-Smirnov should work but no implementation available, critical values of test statistics unknown, impractical in higher dimensions

Kolmogorov-Smirnov test



Here, K-S will be used as benchmark

- one-sample: $D_n = \sup_x |F_n(x) - F(x)|$
- two-sample: $D_{n,m} = \sup_x |F_{1,n}(x) - F_{2,m}(x)|$

One sample TopoTests

Input: sample $X = \{x_1, x_2, \ldots, x_n\}, x_i \in R^d$ and CDF $F : R^d \to [0,1]$.

Step 1: $E_F(\chi(n,r))$, the Blueprint of F

- draw n-element samples $X'_1, X'_2, \ldots, X'_M$ from F
- for each sample X'_i compute its ECC $\chi(C_r(X'_i))$
- $$\frac{1}{M}\sum_{i=1}^{M}\chi(C_r(X'_i)) \xrightarrow[M\to\infty]{a.s.} E_F(\chi(n,r))$$

One sample TopoTests

Input: sample $X = \{x_1, x_2, \dots, x_n\}, x_i \in R^d$ and CDF $F : R^d \to [0, 1]$.

Step 2: variation form $E_F(\chi(n, r))$

- draw a new set of m-element samples $Y'_1, Y'_2, \dots, Y'_m$ from F
- Calculate sup distance between $\chi(C_r(Y'_i)), i = 1, \dots, m$ and average ECC
- determine the threshold value t_α as a $(1 - \alpha)$'th quantile of $\{d_i\}_{i=1}^m$, where α is required level of statistical significance

TopoTests

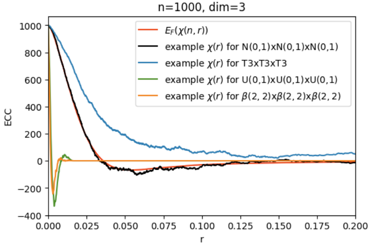
Input: sample $X = \{x_1, x_2, \ldots, x_n\}, x_i \in R^d$ and CDF $F : R^d \to [0, 1]$.

Step 3: Actual testing

- compute the ECC for sample data X: $\chi(C_r(X))$
- compute the l_∞ between $\chi(C_r(X))$ and $E_F(\chi(n, r))$

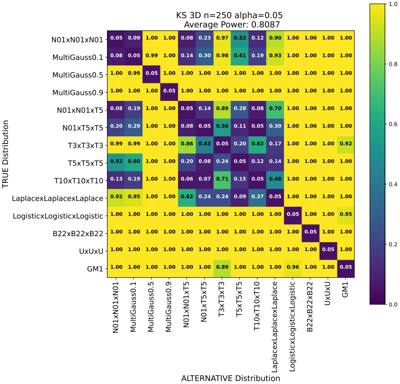
$$D = \sup_{r \in \mathbb{R}} |\chi(C_r(X)) - E_F(\chi(n, r))|$$

- reject H_0 if $D > t_\alpha$
- it is possible to get p-value as well



For the two-sample problem the procedure is slightly different but the idea remains.

Simulation results (one-sample)



average power at $\alpha = 0.05$:

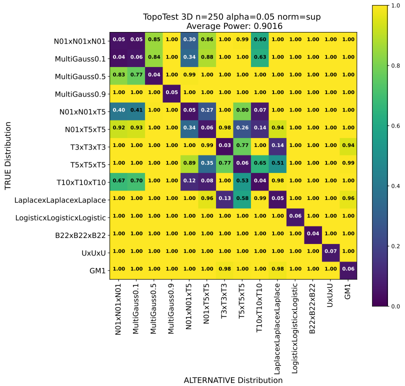
$d = 3, n = 250$ TT:0.9016, KS : 0.8087

$d = 5, n = 500$ TT:0.8465, KS : $---$

Test Power: probability that H_0 is correctly rejected when H_1 is true

- Samples sizes 100–5000 data points
- test power estimated using 1000 MC replications
- power compared with KS ($d \leq 3$)
- α on diagonal is expected
- TopoTests yielded higher power than KS in most of the cases

Simulation results (one-sample)



average power at $\alpha = 0.05$:

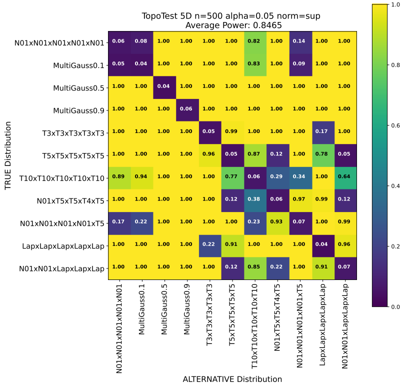
$d = 3, n = 250$ TT:0.9016, KS : 0.8087

$d = 5, n = 500$ TT:0.8465, KS : $---$

Test Power: probability that H_0 is correctly rejected when H_1 is true

- Samples sizes 100–5000 data points
- test power estimated using 1000 MC replications
- power compared with KS ($d \leq 3$)
- α on diagonal is expected
- TopoTests yielded higher power than KS in most of the cases

Simulation results (one-sample)



average power at $\alpha = 0.05$:

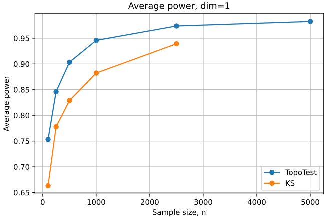
$d = 3, n = 250$ TT:0.9016, KS : 0.8087

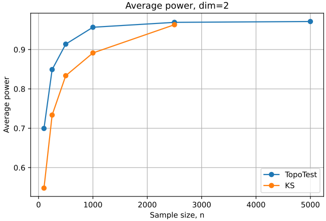
$d = 5, n = 500$ TT:0.8465, KS : $- - -$

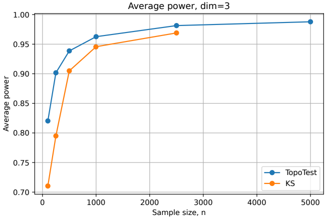
Test Power: probability that H_0 is correctly rejected when H_1 is true

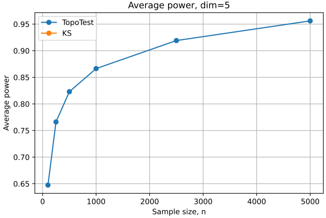
- Samples sizes 100–5000 data points
- test power estimated using 1000 MC replications
- power compared with KS ($d \leq 3$)
- α on diagonal is expected
- TopoTests yielded higher power than KS in most of the cases

Simulation results (one-sample)









So what? Do I really need to know the cumulative distribution function?

Damage identification with TopoTests

Problem statement

Alpha stable noise: intact machine

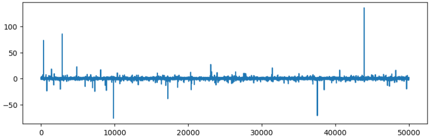
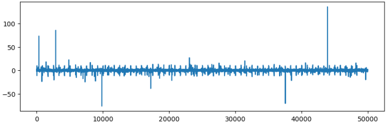


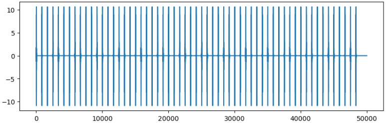
Figure: Crusher, source: Wikipedia

VS.

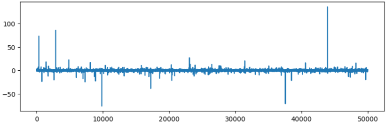
Alpha stable noise + cyclic impulses: malfunctioning machine



=



+



Pipeline

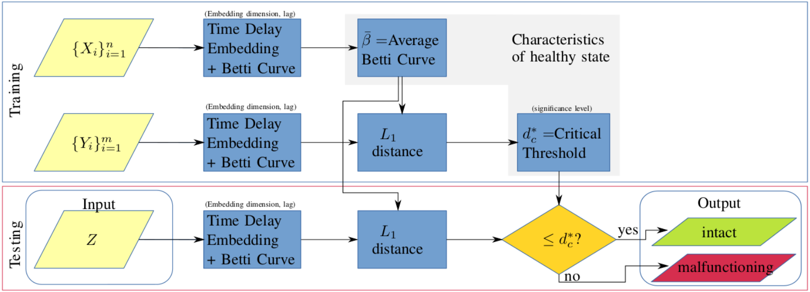


Figure: Flow chart of our testing procedure.

Results: simulated data

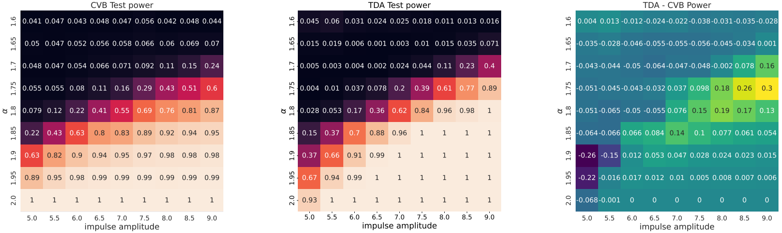


Figure: Comparison state of the art (conditional variance band selector, left) our approach (first Betti curve, middle), and their difference (right). High test power means low frequency of identifying an actually faulty machine as intact.

Results: lab measurement (test bench)

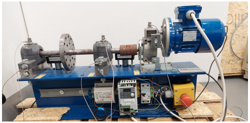


Figure: The test bench.

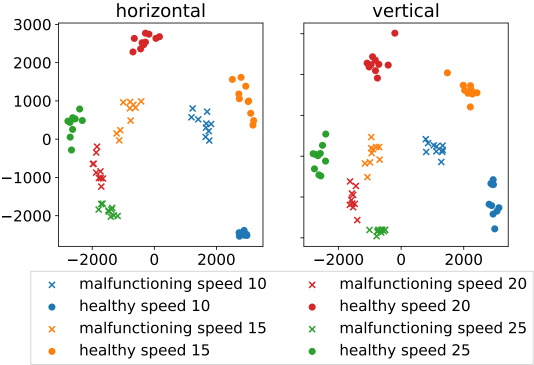


Figure: PCA from Betti curves.

Results: real world measurements (idler)

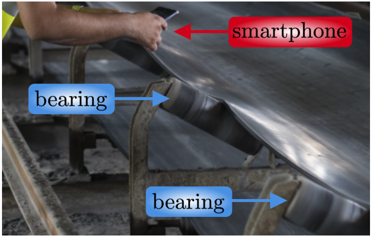


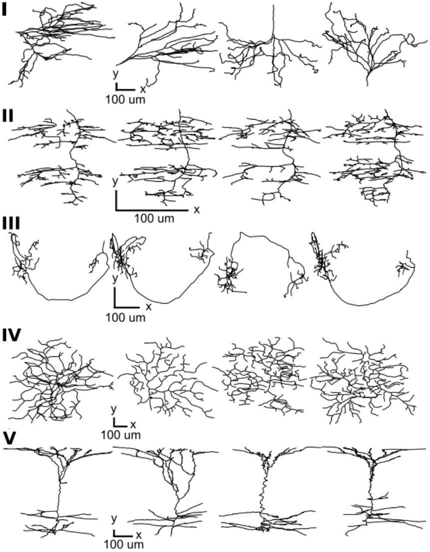
Figure: The idler.

	Industrial data
CVB	87.6 ± 7.16
TDA	92.0 ± 5.61
CVB +TDA	96.1 ± 4.50

Table: Mean accuracy of SVM classifier [%] and standard deviation.

Shapes of neurons (and trees, and graphs)

Shape → function

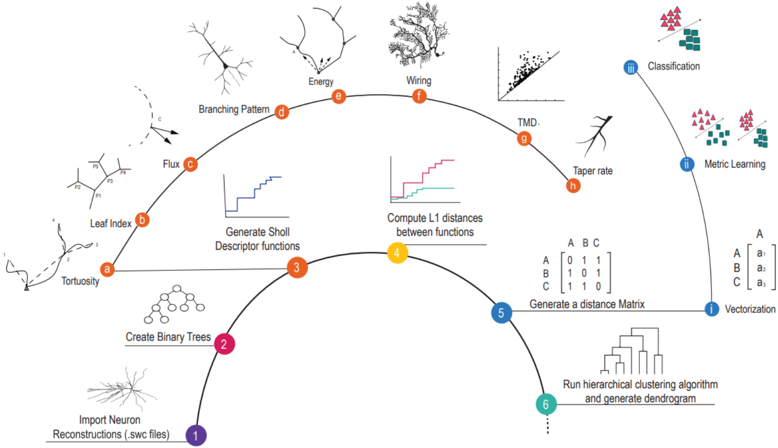


(I) cat, (II) dragonfly, (III) fruit fly, (IV) mouse and (V) rat

Shapes of rooted trees in $\mathbb{R}^3$

- Neurons are particular instance of trees in $\mathbb{R}^3$.
- Root is the soma.
- Morphological descriptors : number of leafs, total occupied volume, polarity, ... (classical)
- Sholification of morphological descriptors (with Khalil, Kallel, Farhad) – descriptor as a function of distance from the somma.
- Branching structure of a tree – mergegrams, TMD and other invariants.

Sholl descriptor



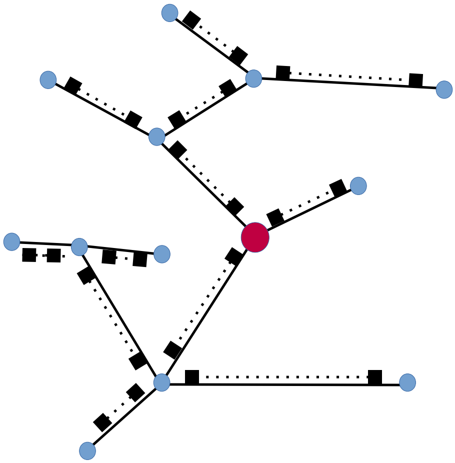
Invriant as a function of distance from soma

How to get a descriptor of a shape of a tree?

- Cut the tree into branches
- Compute invariants of branches

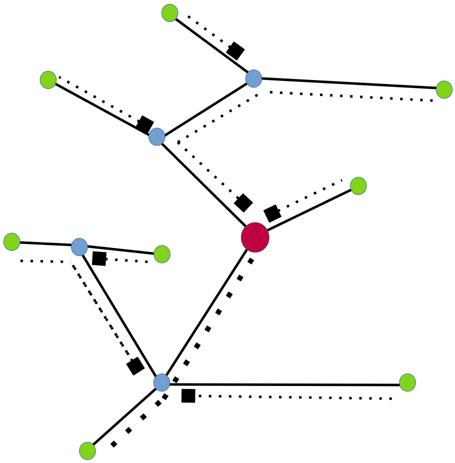
Mergegram branch decomposition

Cut all branching nodes.

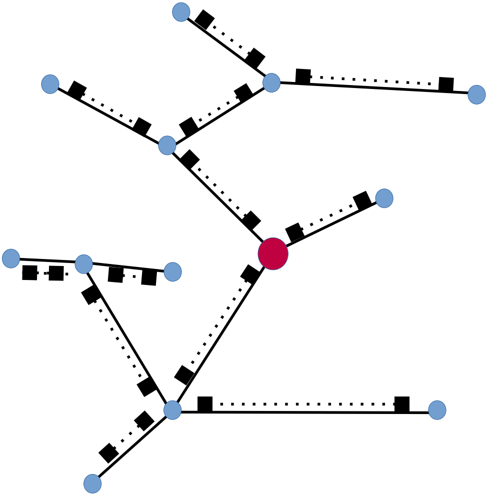


TMD Branch decomposition

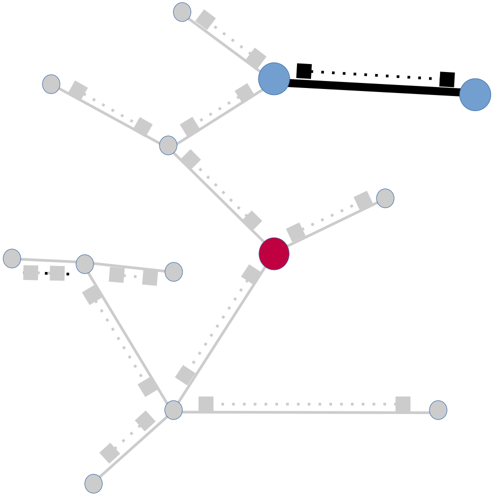
Let the longest branch to continue towards the root



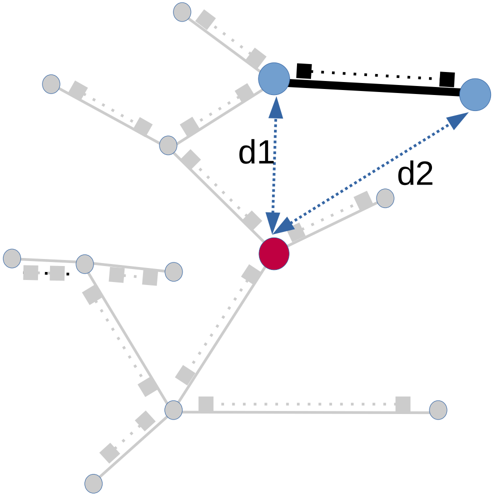
From branches to diagrams



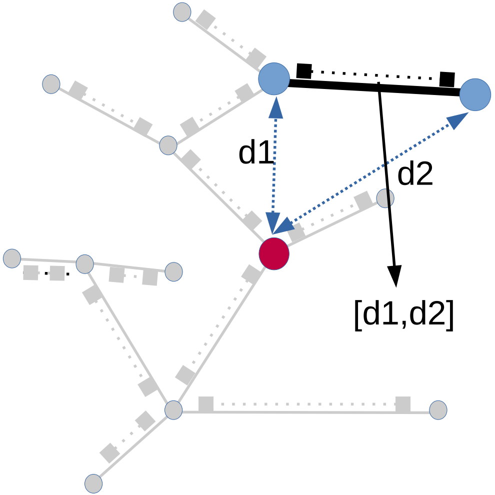
From branches to diagrams



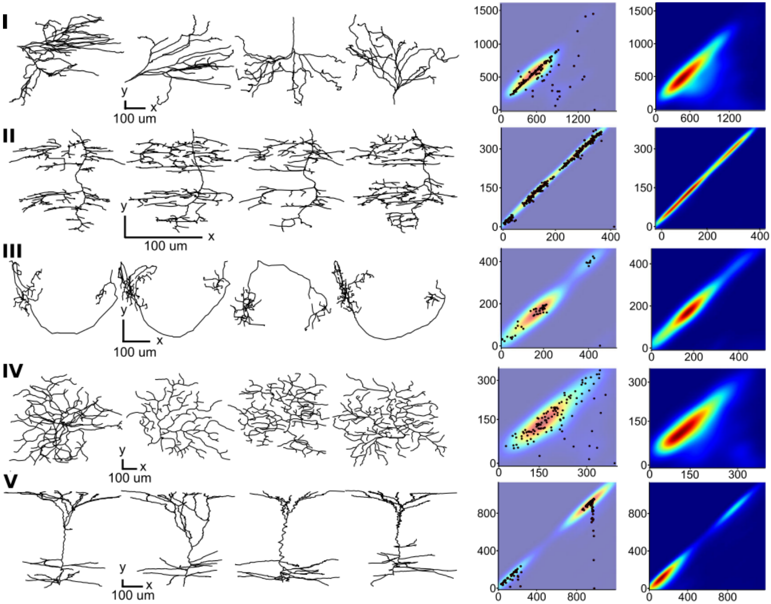
From branches to diagrams



From branches to diagrams



TMD descriptors of trees



Is a single descriptor sufficient?

- Variety of tree structures is huge,
- Each descriptor is capturing a single aspect of it.
- Not sufficient to capture the complexity of possible trees.
- Solution: Combine different descriptors into a single one.

Multiple descriptors for labeled data

- For labeled data, combine them into single distance

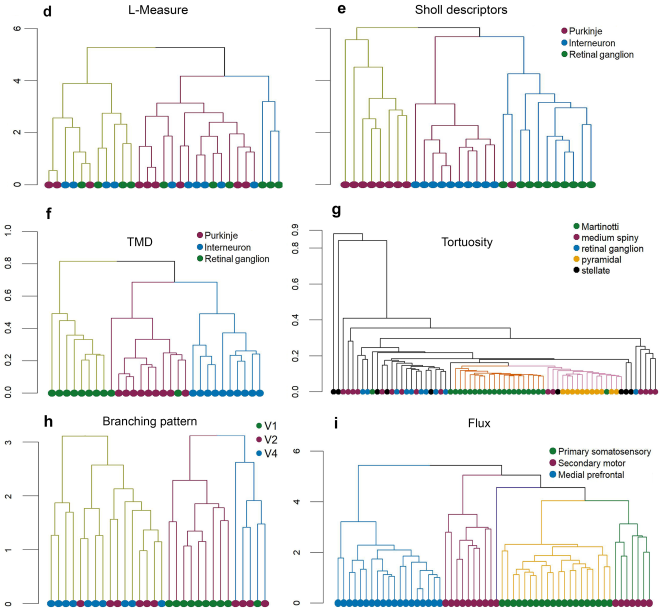
$$d = \alpha_1 d_1 + \alpha_2 d_2 + \ldots + \alpha_n d_n$$

  and optimize $\alpha_1, \ldots, \alpha_n$ for best separation,
- Use Metric Learning and Mahalanobis distances

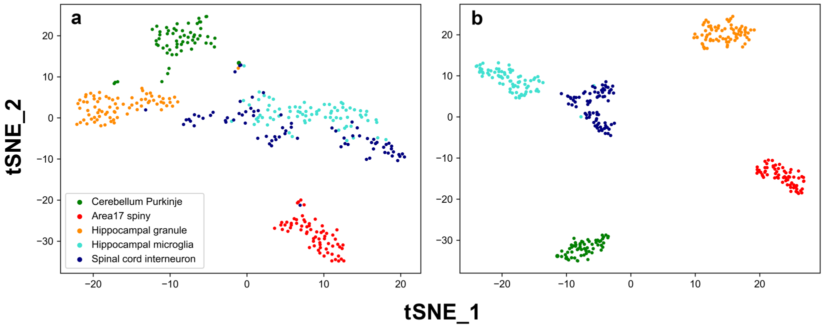
$$D(x, y) = \sqrt{(Lx - Ly)^T (Lx - Ly)}$$

  to obtain best separation.

Some classification results



Some classification results



Euclidean vs Metric Learning–transformed space.

Wrap up

- Data points, images, physical phenomena often have some intrinsic shape,
- Understanding this shape is important to understand the underlying process,
- Topological data analysis provides tools to understand the shape of data.

The TDA-Team

Dioscuri Centre in Topological Data Analysis

Thank you for your time

Dioscuri Centre in Topological Data Analysis

DIOSCURI
CENTRE IN TOPOLOGICAL DATA ANALYSIS

MAX PLANCK GESELLSCHAFT
NATIONAL SCIENCE CENTRE POLAND

Jointly sponsored by
Ministry of Science and Higher Education Republic of Poland
Federal Ministry of Education and Research

Paweł Dłotko
pdlotko @ impan.pl
pdlotko @ gmail
pawel_dlotko @ skype