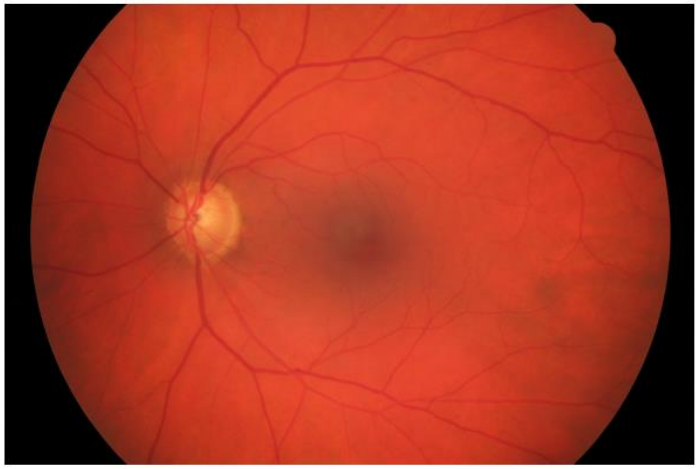
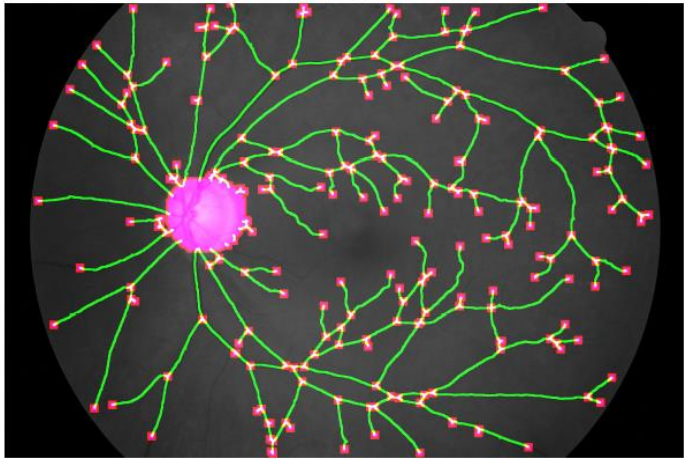




Complex network approach to high dimensional data analysis

Cristina Masoller
Departamento de Fisica,
Universitat Politècnica de Catalunya

Conference on Complex Systems 2023
Salvador, Bahia, Brazil

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

Campus d'Excel·lència Internacional

ICREA

AGAUR

GOBIERNO DE ESPAÑA
MINISTERIO DE CIENCIA E INNOVACIÓN

CCS

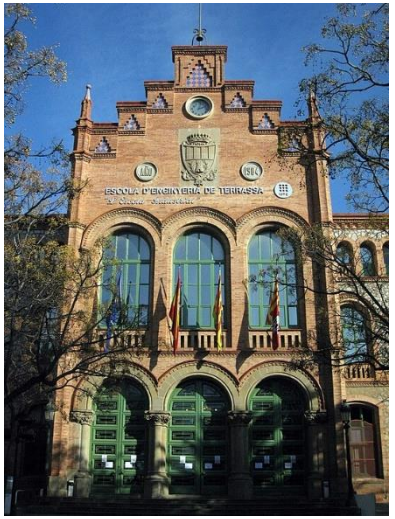
Presentation

- Originally from Montevideo, Uruguay.
- At Universitat Politecnica de Catalunya since 2004, full professor since 2018
- Research group: Dynamics, Nonlinear Optics and Lasers

www.donll.upc. edu

cristina.masoller@upc.edu @cristinamasoll1

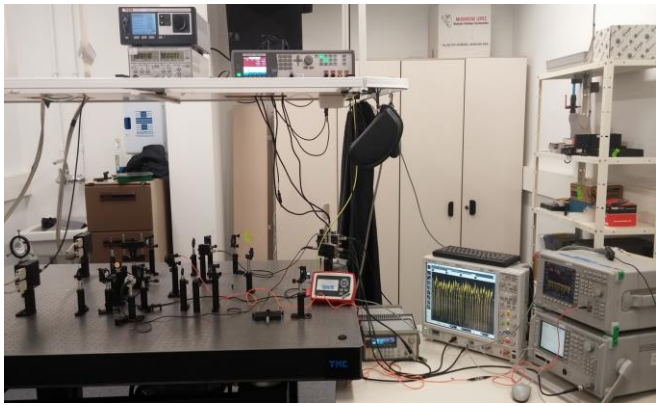
Where are we? UPC Campus Terrassa



School of Industrial, Aerospace & Audiovisual Engineering of Terrassa


Viernes, 25 de septiembre de 2009 Diari de Terrassa

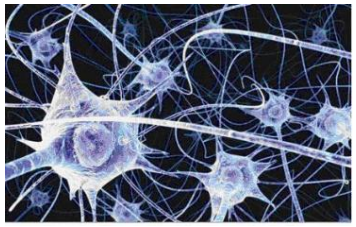
El edificio Gaia centraliza grupos científicos consolidados y emergentes.



Laser lab

cristina.masoller@upc.edu @cristinamasoll1

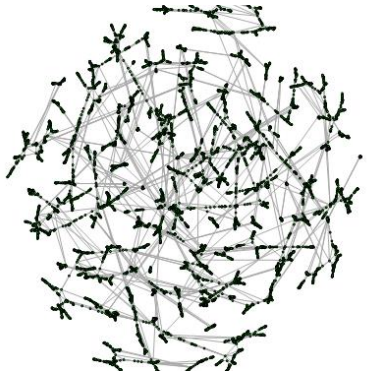
Research lines

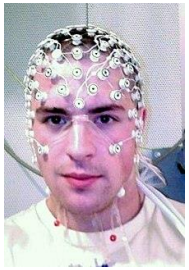


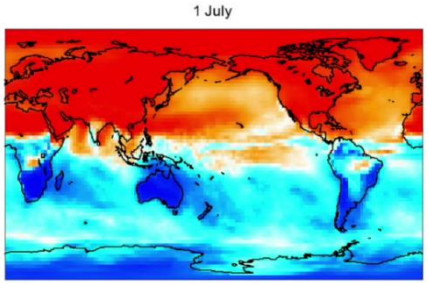
Nonlinear dynamics and complex systems

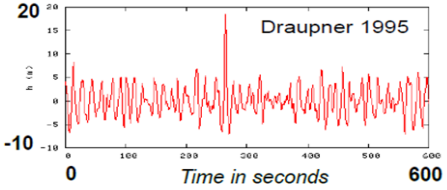
Data analysis techniques

Applications









https://shiny.rcg.sfu.ca/u/rdmorin/scholar_googler/

cristina.masoller@upc.edu @cristinamasoll1

Outline

- Network-based tools for retina fundus image analysis
- Anomaly / outlier detection in high dimensional data
- Ordinal (symbolic) analysis of vegetation images
- Take home messages

cristina.masoller@upc.edu @cristinamasoll1

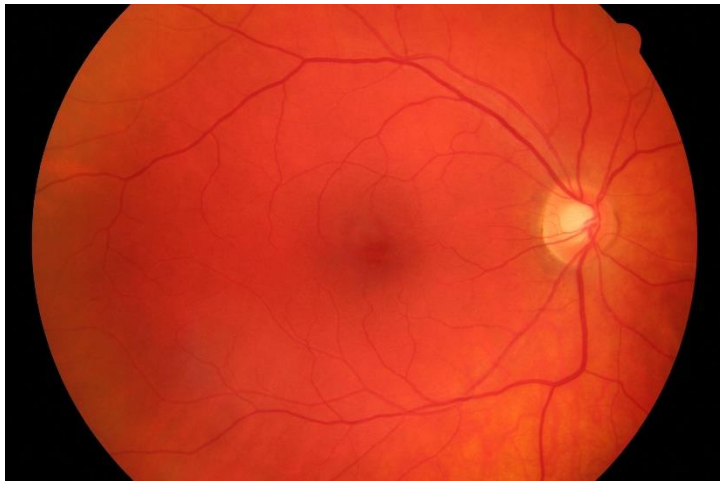
Motivation

- The analysis of retina fundus images allows for the diagnosis of eye diseases (glaucoma, diabetes) & follow up of treatments.
- Biometric identity identification.
- Opportunity to detect other diseases (alterations in retina network may reflect alterations in other arterial systems).

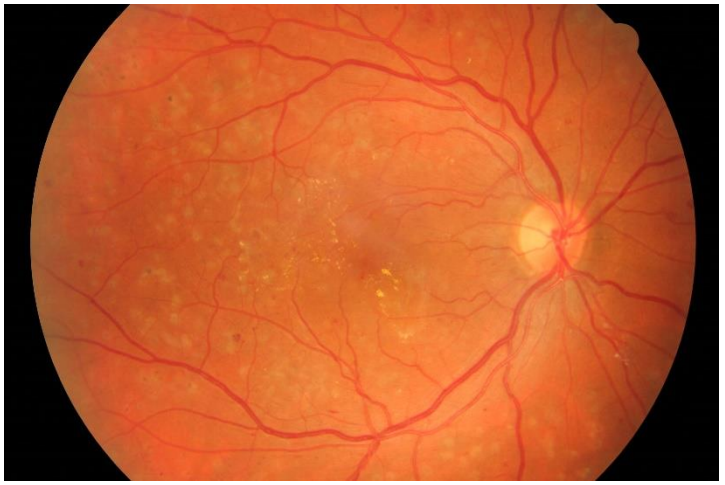
cristina.masoller@upc.edu @cristinamasoll1

Examples of retina fundus photos

Healthy



Diabetic

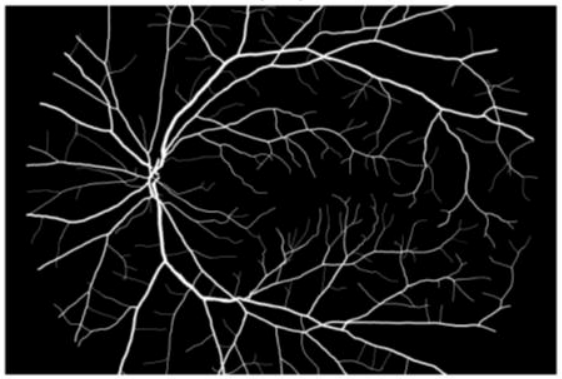


cristina.masoller@upc.edu @cristinamasoll1

Image data base

- High-resolution (3504 × 2336 pixels) with
  - 15 healthy subjects
  - 15 glaucoma
  - 15 diabetic retinopathy

- For every subject there is
  - fundus photography
  - manual segmentation of the vessels done by an expert.



https://www5.cs.fau.de/research/data/fundus-images/

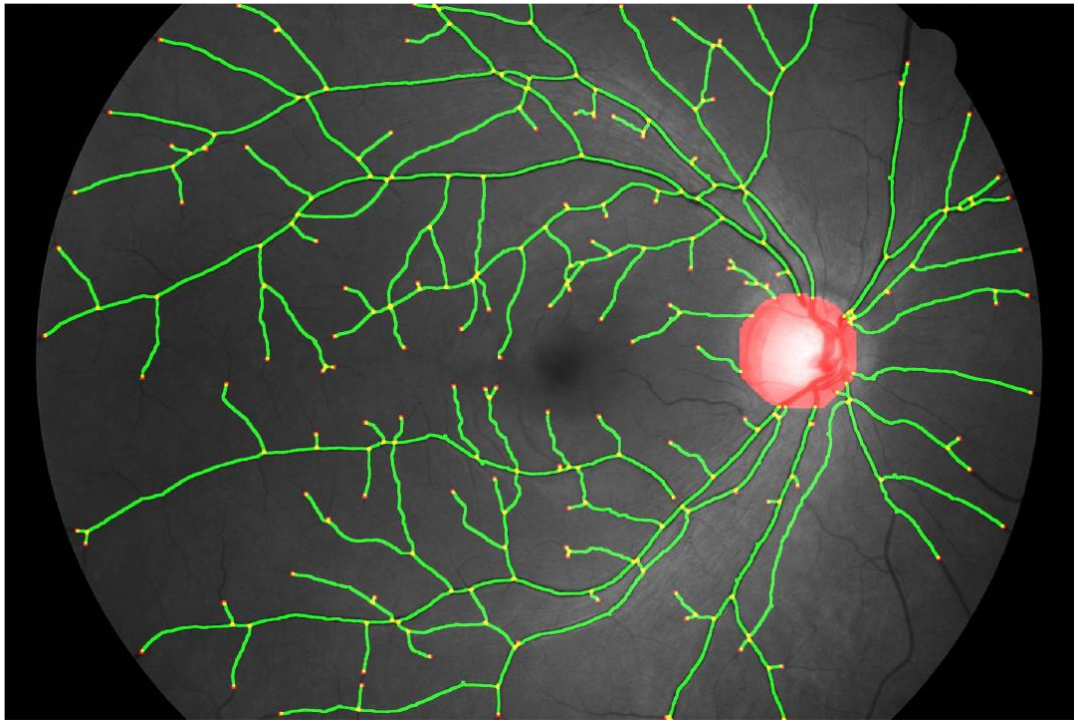
cristina.masoller@upc.edu @cristinamasoll1

Image analysis steps

1. Pre-process images

(for automatic segmentation)

2. Extract networks
3. Compare networks

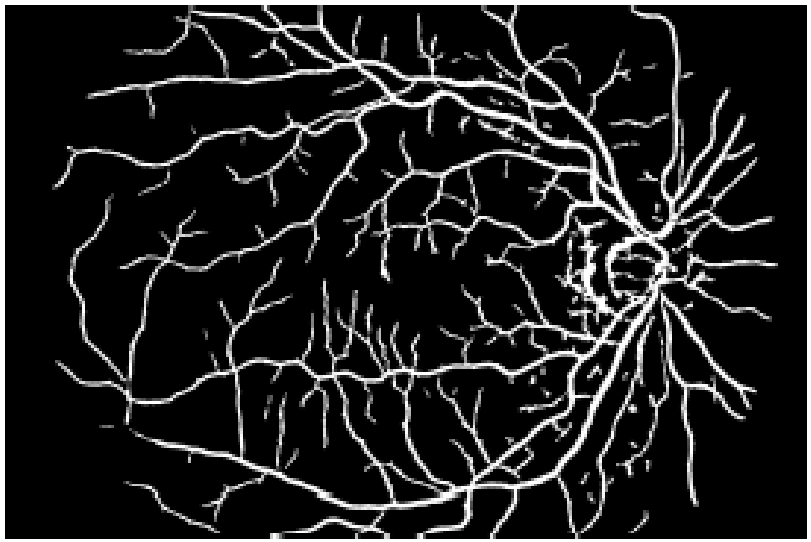


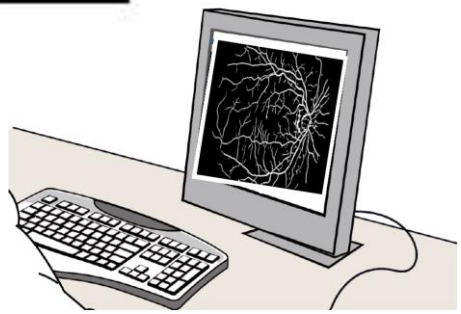
P. Amil et al, Network-based features for retinal fundus vessel structure analysis, PLoS ONE 14 e0220132 (2019).

cristina.masoller@upc.edu @cristinamasoll1

Step 1: Unsupervised algorithm for pre-processing the images

Automated segmentation

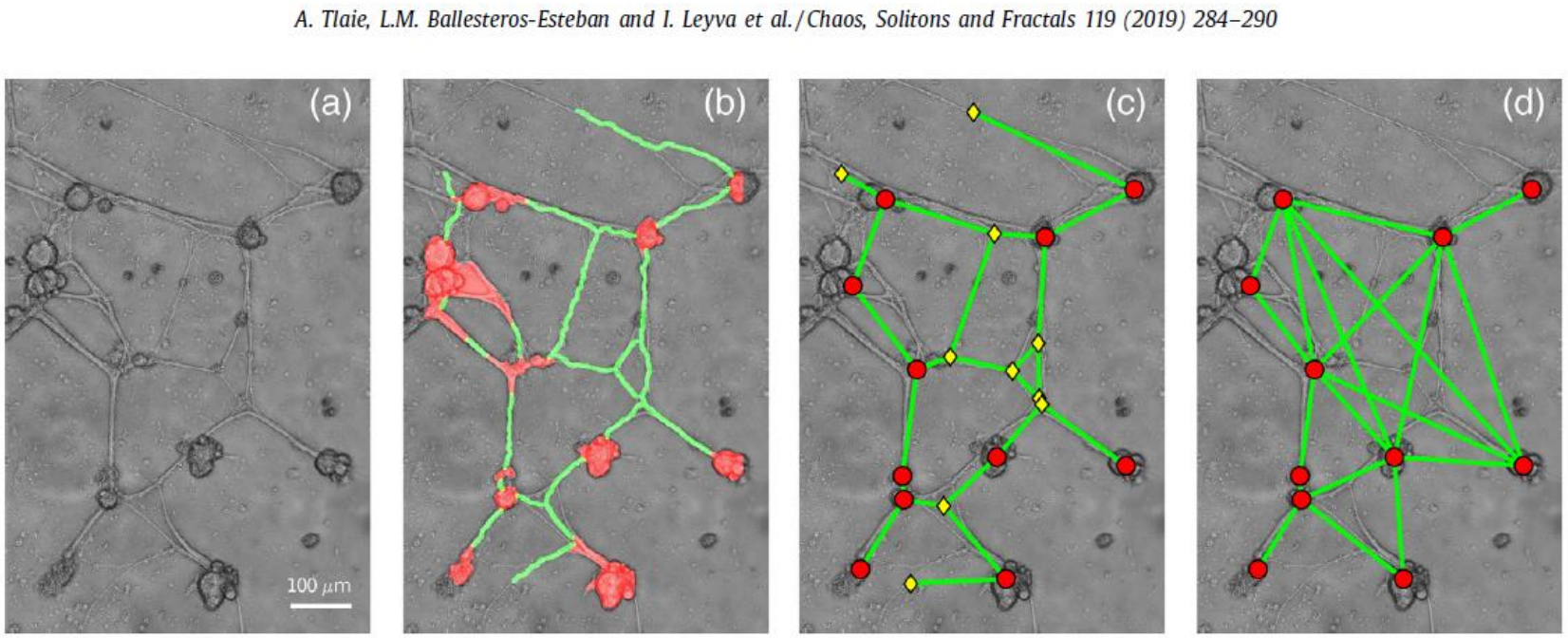




P. Amil et al, Network-based features for retinal fundus vessel structure analysis, PLoS ONE 14 e0220132 (2019).

cristina.masoller@upc.edu @cristinamasoll1

Segmentation algorithm adapted from an algorithm developed to segment images of cultured neural networks.

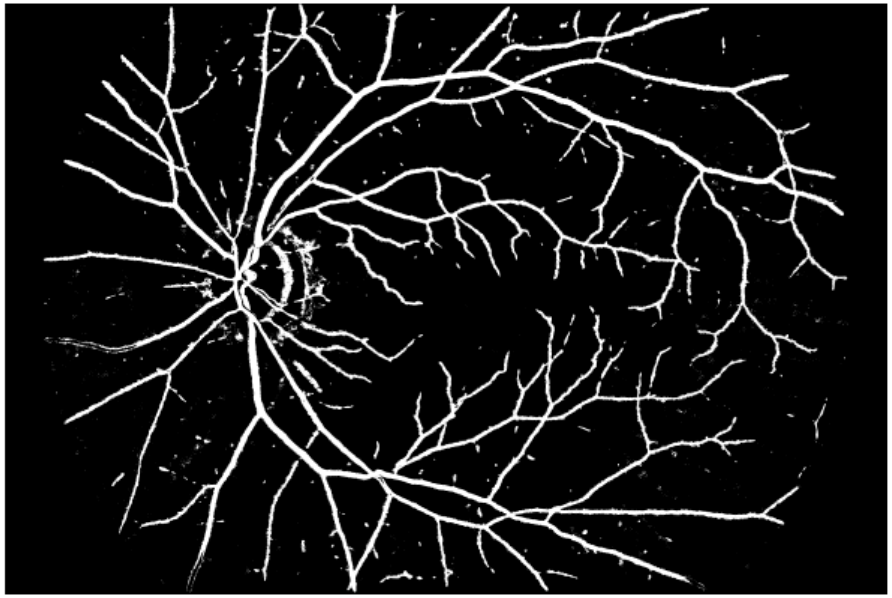


Santos-Sierra D, Sendiña-Nadal I, Leyva I, et al. Cytometry Part A. 87, 513 (2015).

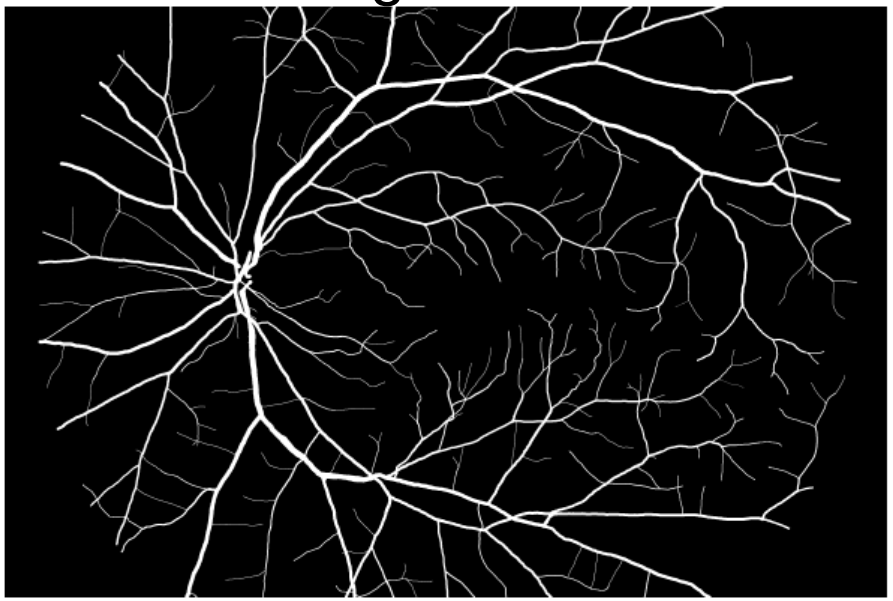
cristina.masoller@upc.edu @cristinamasoll1

Comparison

Automated segmentation



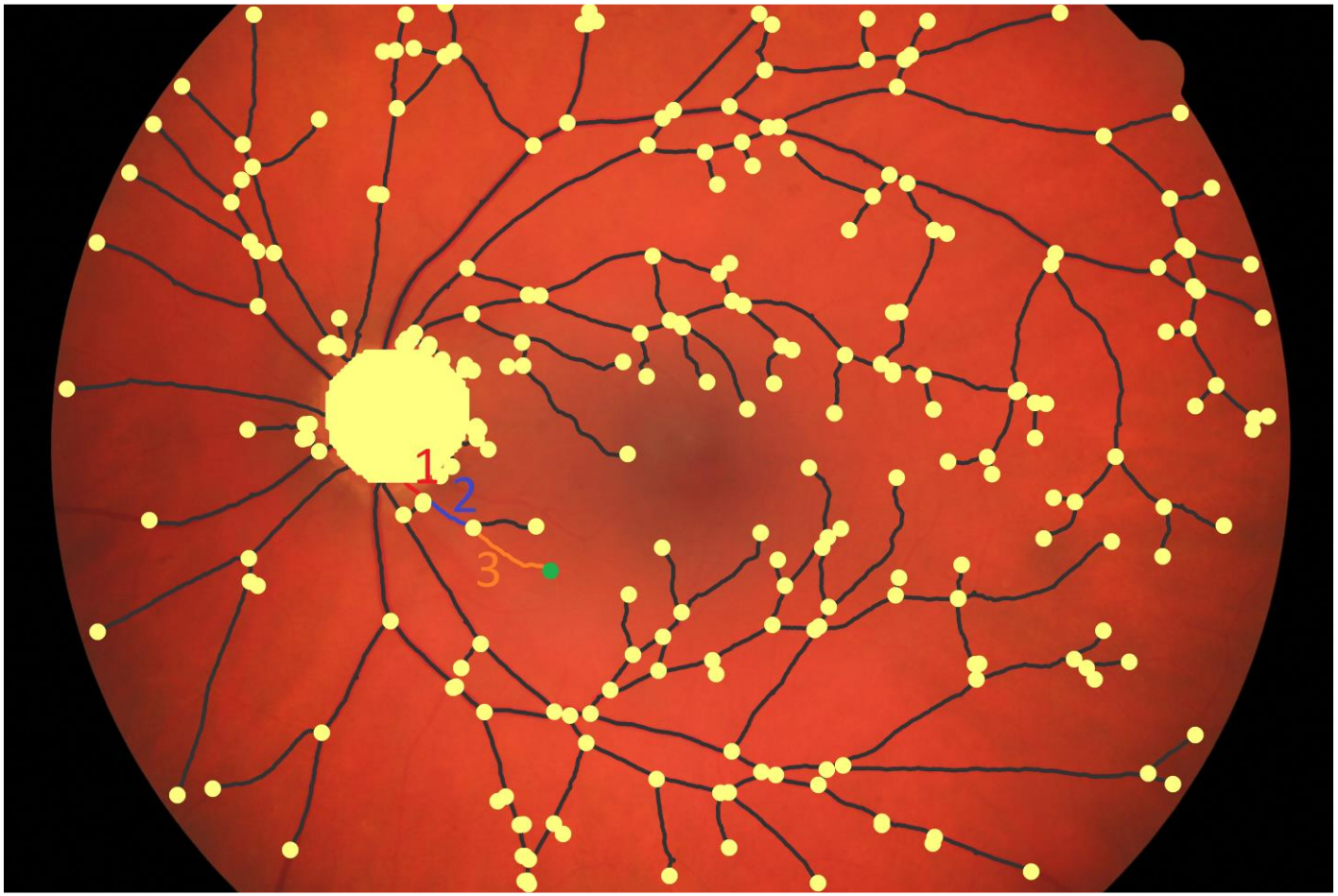
Manual segmentation



P. Amil et al, Network-based features for retinal fundus vessel structure analysis, PLoS ONE 14 e0220132 (2019).

cristina.masoller@upc.edu @cristinamasoll1

Step 2a: Extract network (identification of nodes and links).



P. Amil et al, Network-based features for retinal fundus vessel structure analysis, PLoS ONE 14 e0220132 (2019).

cristina.masoller@upc.edu @cristinamasoll1

Step 2b: Define the weights of the links

$$w_{i,j} = (L_{i,j})^l \, (W_{i,j})^a$$

length and width (in # of pixels) of the segment that connects nodes i and j

- For diabetic retinopathy (DR) length/area (l = 1, a = -2) provide the best differentiation between groups (DR produces neovascularization, which may affect the vessels' flow capacity).

- For glaucoma patients, the volume performed the best (glaucoma is linked to an increase of the intraocular pressure, which perhaps modifies the volume of the vessels).

P. Amil et al, Network-based features for retinal fundus vessel structure analysis, PLoS ONE 14 e0220132 (2019).

cristina.masoller@upc.edu @cristinamasoll1

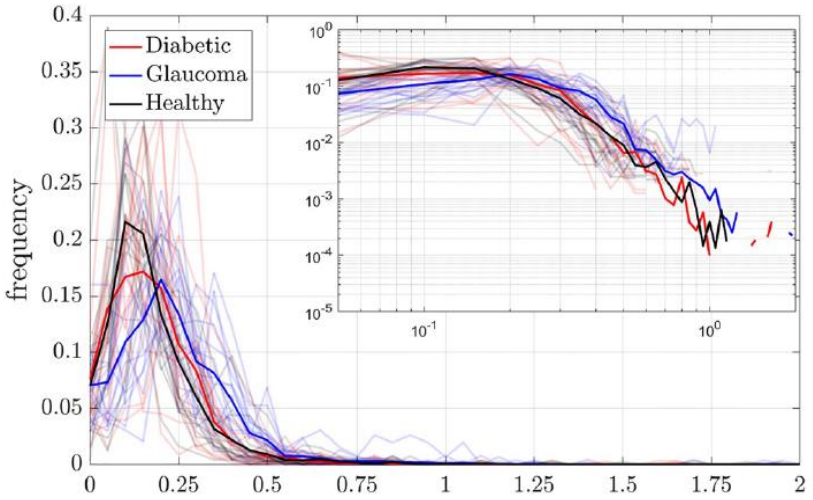
Step 3: Pair-wise network comparison

How to compare networks that have *different number of nodes*?

Our approach: Extract distributions from each network & compare the distributions.

- Distribution of shortest paths to the central node
- Distribution of average weights along the shortest path to the central node
- Weighted degree distribution

Example: Distribution of average weights in shortest paths, extracted from the automatic segmentation.



Average weight (l=1, a=-2)

cristina.masoller@upc.edu @cristinamasoll1

Step 3: Pair-wise comparison of the distributions

- How to compare two distributions? Many options!

  S.-H. Cha, Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions, Int. J of. Math. Models and Meth. 1, 300 (2007).

- The Jensen-Shannon (JS) divergence:

$$D_{ij} = JS[P_i, P_j] = H[(P_i + P_j)/2] - H[P_i]/2 - H[P_j]/2,$$

  (*H*: entropy of the distribution *P*)

- For each image "i" we obtain a "vector of distances" to all the other images

  $D_i = \{D_{i1}, D_{i2}, \ldots, D_{iN}\}$ (N = number of images = 45, $D_{ii}=0$)

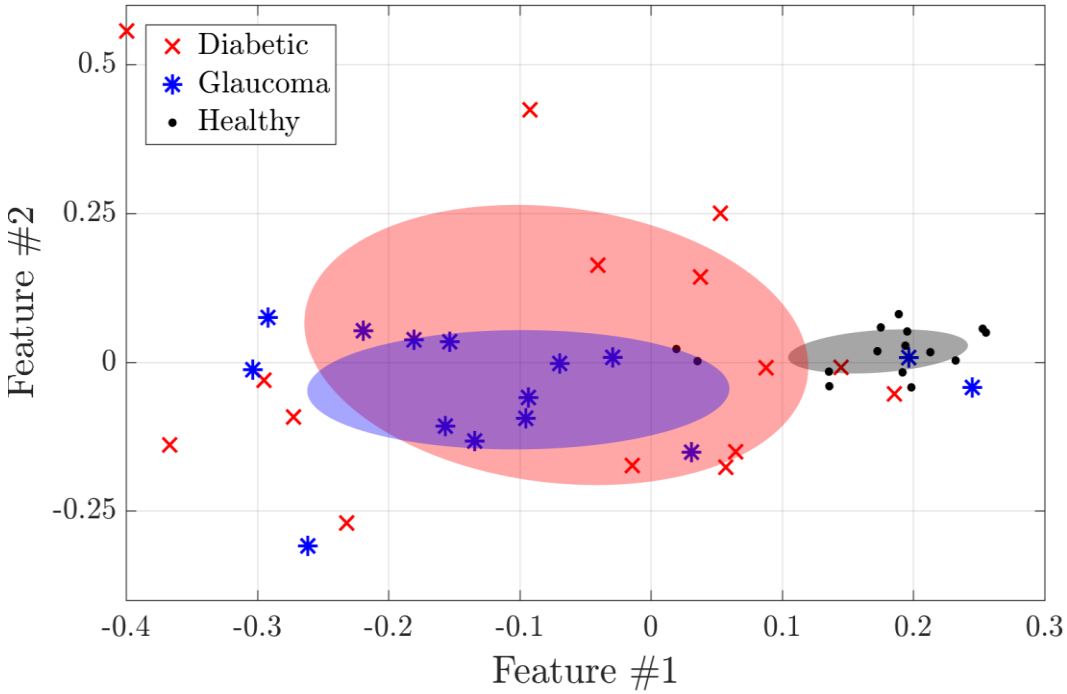
- This vector has 45 "features" that characterize the i-th image (3504 × 2336 pixels) in relation to the other images.

cristina.masoller@upc.edu @cristinamasoll1

Step 5: Apply a nonlinear dimensionality reduction algorithm (IsoMap) to obtain only 2 features for each image.

J. B. Tenenbaum et al., Science 290, 2319 (2000).

From the distribution of shortest path distances to the central node, in the manual segmentation:

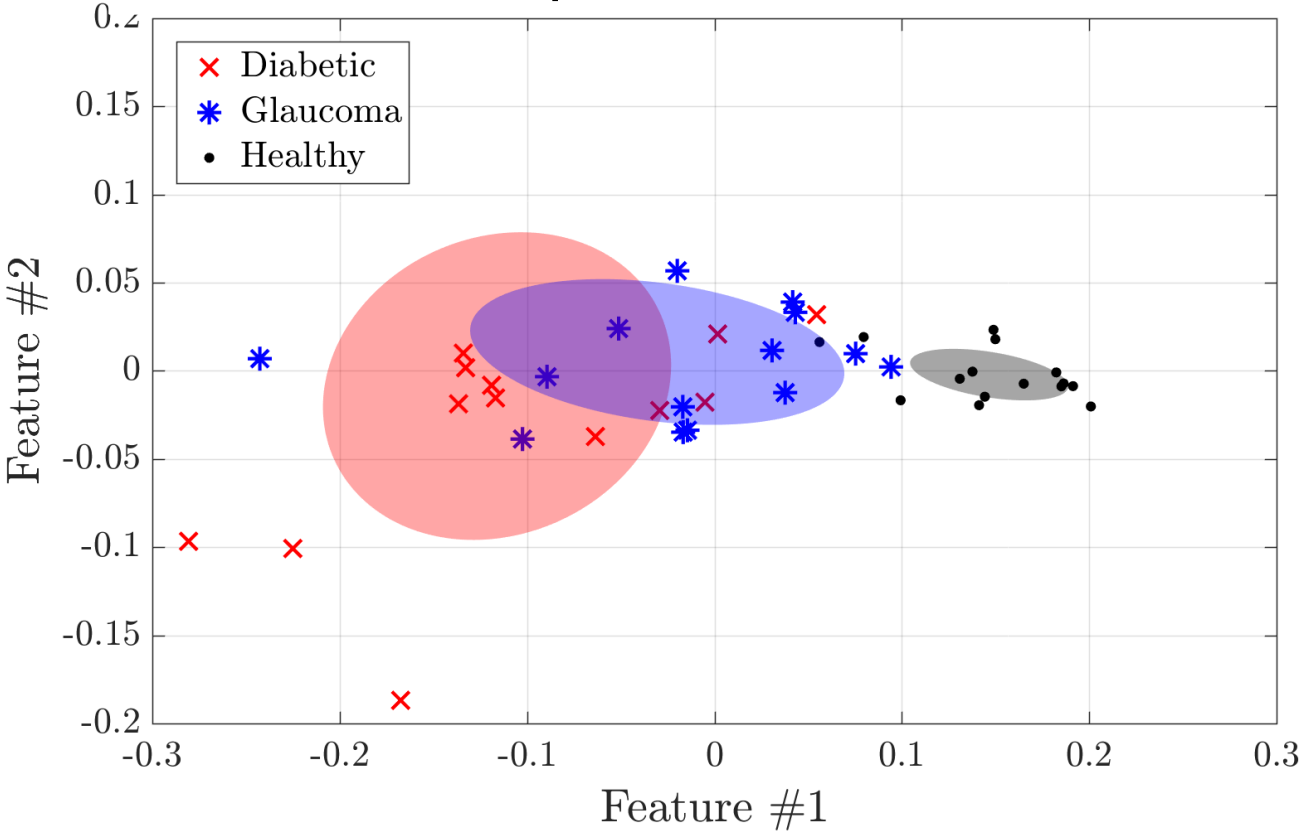


P. Amil et al, Network-based features for retinal fundus vessel structure analysis, PLoS ONE 14 e0220132 (2019).

cristina.masoller@upc.edu @cristinamasoll1

Performance of another network-based feature in the manual segmentation:

Distribution of average weights along the shortest path to central node

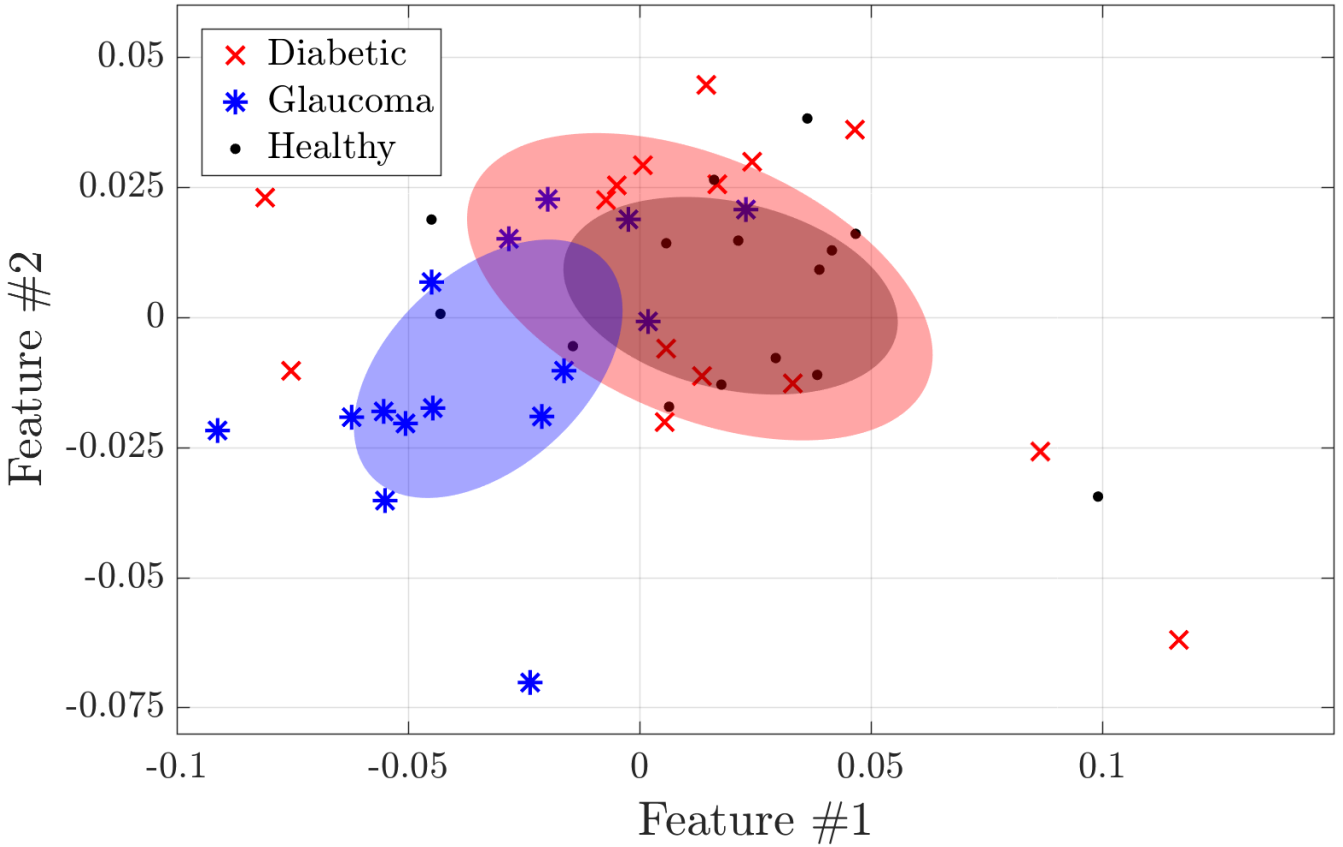


P. Amil et al, Network-based features for retinal fundus vessel structure analysis, PLoS ONE 14 e0220132 (2019).

cristina.masoller@upc.edu @cristinamasoll1

In the automated segmentation

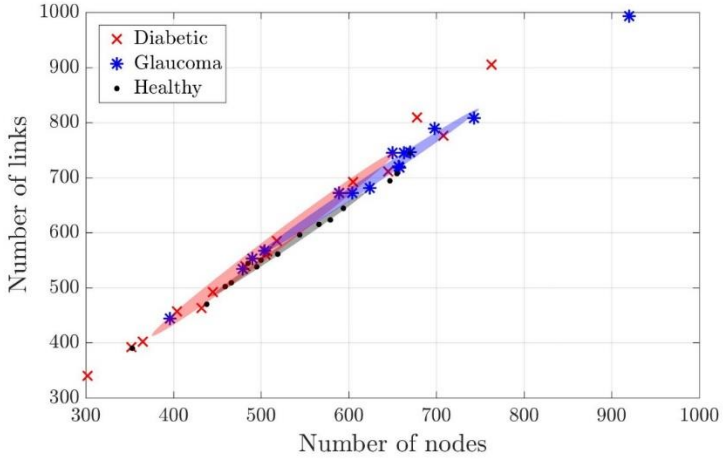
The distribution of average weights along the shortest path to the central node separates glaucoma:



P. Amil et al, Network-based features for retinal fundus vessel structure analysis, PLoS ONE 14 e0220132 (2019).

cristina.masoller@upc.edu @cristinamasoll1

Simple network features do not differentiate



RESEARCH ARTICLE

Network-based features for retinal fundus vessel structure analysis

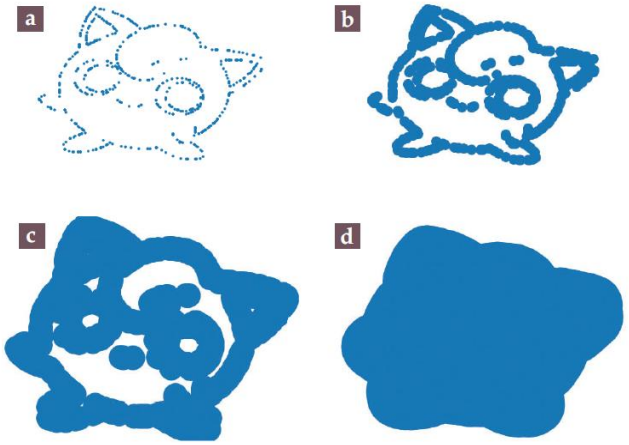
Pablo Amil[1]*, Cesar F. Reyes-Manzano[2], Lev Guzmán-Vargas[2], Irene Sendiña-Nadal[3,4], Cristina Masoller[1]

1 Nonlinear Dynamics, Nonlinear Optics and Lasers, Universitat Politècnica de Catalunya, Terrassa, Spain, 2 Unidad Profesional Interdisciplinaria en Ingeniería y Tecnologías Avanzadas, Instituto Politécnico Nacional, Gustavo A. Madero, Ciudad de México, México, 3 Complex Systems Group & GISC, Universidad Rey Juan Carlos, Madrid, Spain, 4 Center for Biomedical Technology, Universidad Politécnica de Madrid, Madrid, Spain

PLOS ONE | https://doi.org/10.1371/journal.pone.0220132

cristina.masoller@upc.edu cristinamasoll1

We also tried Topological Data Analysis



M. Porter et al, Physics Today, January 2023

RESEARCH ARTICLE

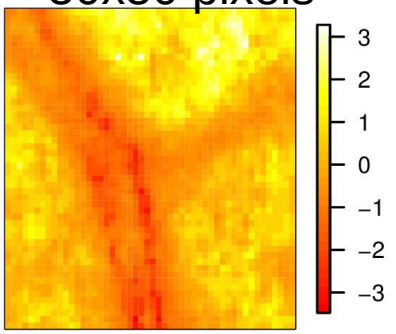
Topological data analysis of high resolution diabetic retinopathy images

Kathryn Garside[1]*, Robin Henderson[1], Irina Makarenko[1], Cristina Masoller[2]

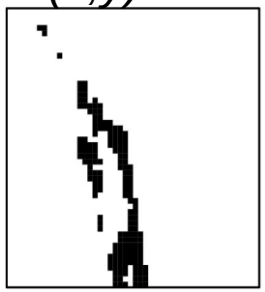
1 School of Mathematics, Statistics and Physics, Newcastle University, Newcastle upon Tyne, United Kingdom, **2** Department of Physics, Universitat Politecnica de Catalunya, Barcelona, Spain

PLOS ONE | https://doi.org/10.1371/journal.pone.0217413

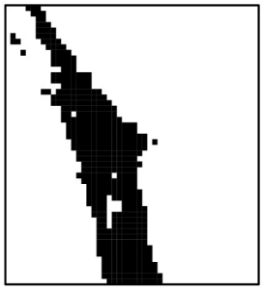
50x50 pixels



z(x,y)<-1.5



z(x,y)<-0.9



cristina.masoller@upc.edu @cristinamasoll1

Outline

- Network-based tools for retina fundus image analysis
- Anomaly / outlier detection in high dimensional data
- Ordinal (symbolic) analysis of vegetation images
- Take home messages

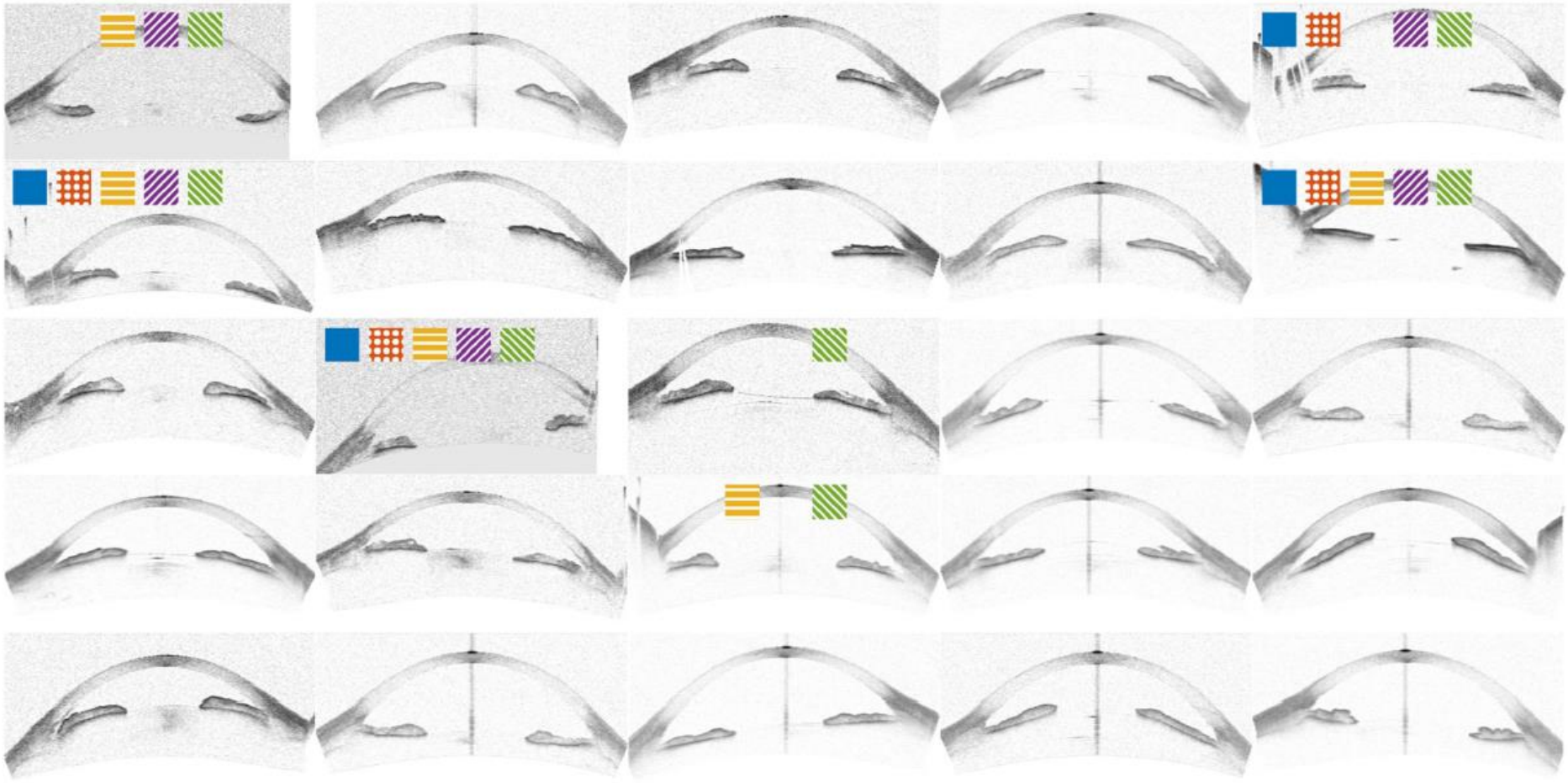
cristina.masoller@upc.edu @cristinamasoll1

Motivation: analysis of OCT anterior chamber images

P. Amil et al.,
Sci. Rep. 9, 1157 (2019).

cristina.masoller@upc.edu @cristinamasoll1

To improve the performance: remove images with artifacts from training data



P. Amil, N. Almeida and C. Masoller, "Outlier mining methods based on graph structure analysis", Front. Phys. 7, 194 (2019).

cristina.masoller@upc.edu @cristinamasoll1

Strange observation/data: an outlier or an anomaly/artifact?

- *Anomaly*: a data point that cannot be explained given current knowledge of the process that generates the data.
- *Outlier*: a "legitimate" data point that is far from the center of the distribution that characterizes the process.
- *Novelty* detection: "new" event/data not seen before.

Types:

- *Point anomalies*: a data point that is anomalous with respect to the rest of the data.
- *Contextual anomalies*: a data point that is anomalous in a specific context.
- *Collective anomalies*: a set of data points that are not anomalies by themselves, but their collective occurrence is anomalous.

V. Chandola et al., ACM Comput. Surveys 41, 15 (2009)

cristina.masoller@upc.edu @cristinamasoll1

Examples of point anomalies

Machine learning “practical” definition: removing anomalies from the training data improves the algorithm’s performance.

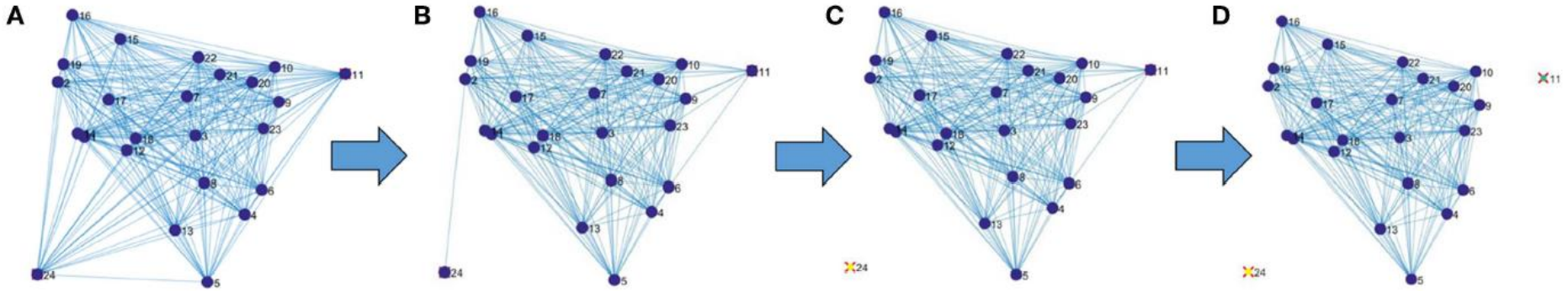
cristina.masoller@upc.edu @cristinamasoll1

We consider high-dimensional datasets where a distance can be defined between pairs of items of the dataset

Feature vectors of items *i* and *j:* $\{f_{i1} \dots f_{iM}\}$ $\{f_{j1} \dots f_{jM}\}$

Euclidian distance: $$d_{ij} = \sqrt{\sum_{k=1}^{M} (f_{ik} - f_{jk})^2}$$

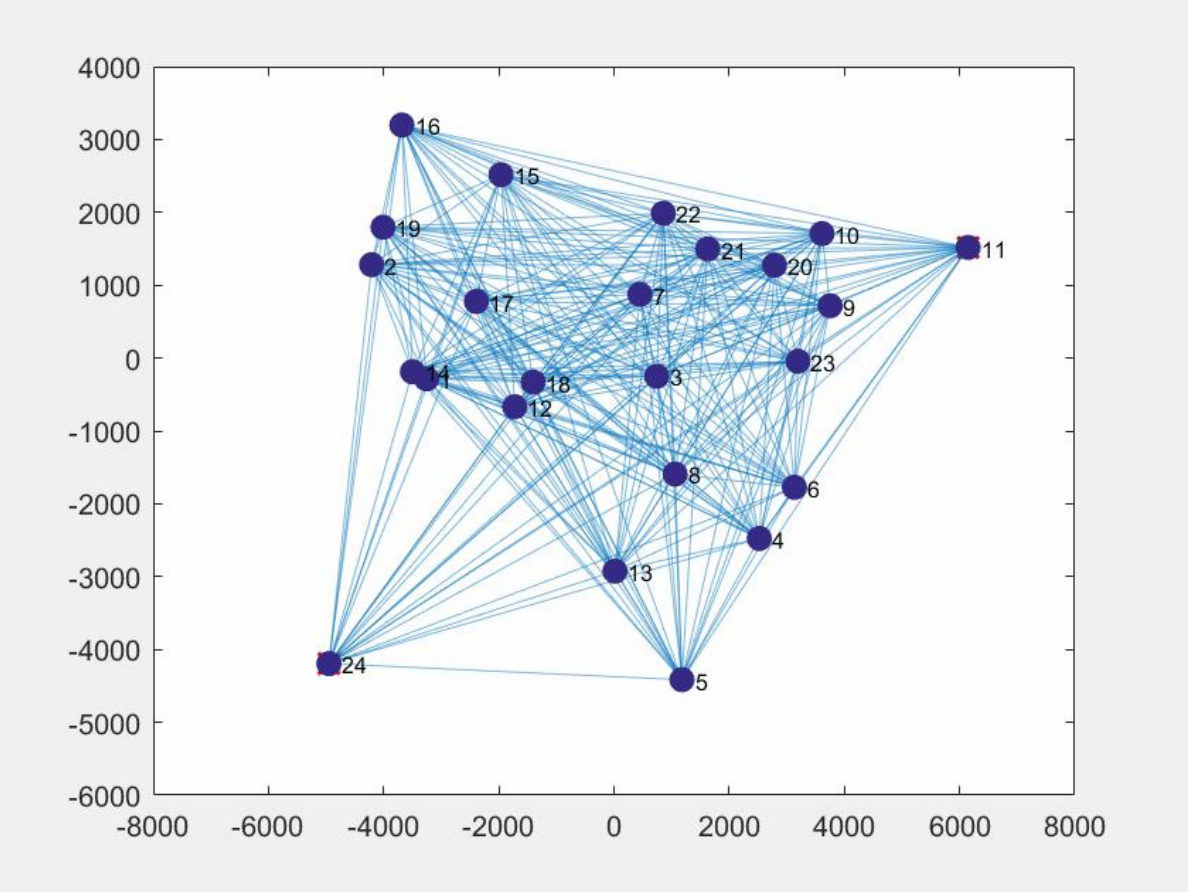
First method: Outlier detection using percolation (OP)



Outlier score = order in which elements disconnect from the giant component. **Parameter free.**

cristina.masoller@upc.edu @cristinamasoll1

First method: Outlier detection using percolation



Outlier score = order in which elements disconnect from the giant component.

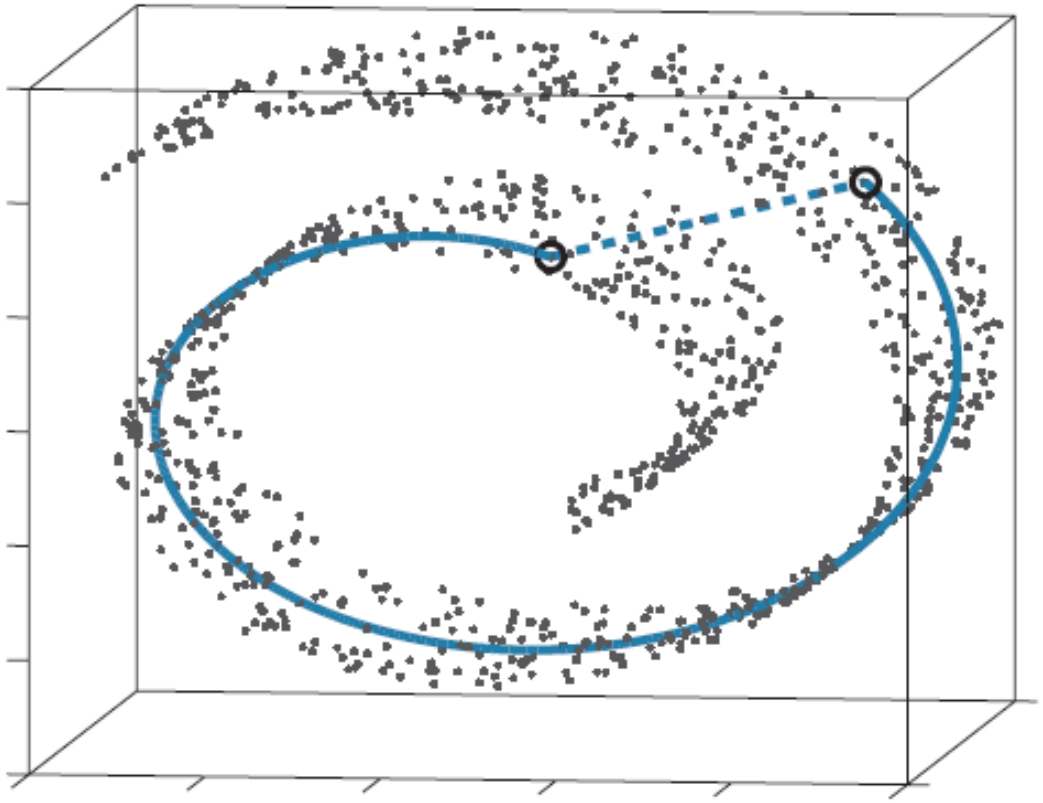
Parameter free.

P. Amil et al., Front. Phys. 7, 194 (2019).

cristina.masoller@upc.edu @cristinamasoll1

Second method: nonlinear dimensionality reduction

- *Main idea: how well or how poorly an element fits in the learned manifold*.



Distance in the high dimensional space (dash) and distance in the learned (lower dimensional) manifold (solid).

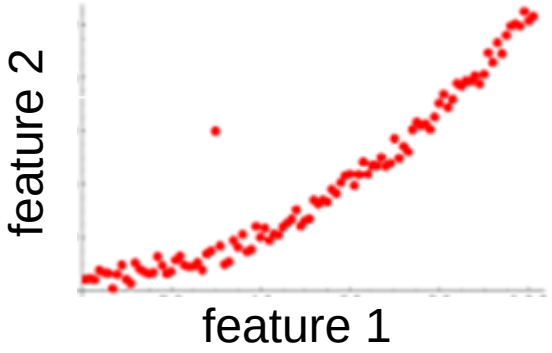
IsoMap, Tenenbaum et al., Science 290, 2319 (2000).
P. Amil, N. Almeida and C. Masoller, Front. Phys. 7, 194 (2019).

cristina.masoller@upc.edu @cristinamasoll1

Steps

- Apply *IsoMap* to the distance matrix $\mathbf{D_{ij}}$ to obtain
  - a new set of features
  - a new distance matrix in the geodesic space, $\mathbf{D^G_{ij}}$
- With the new features, recalculate the distance matrix $\mathbf{D'_{ij}}$
- For each element, calculate correlation, between $\mathbf{D^G_{ij}}$ and $\mathbf{D'_{ij}}$
- Outlier score: $OS_i = 1-\rho_i^2$
- Two parameters (integers):
  - Dimension of reduced space
  - # of geodesic neighbors

Note: we don't use the features returned by IsoMap to assign outlier scores



cristina.masoller@upc.edu @cristinamasoll1

Comparison with other distance-based outlier mining methods

- Distance to center of mass (d2CM): an outlier score is assigned according to the distance of an element to the center of mass.

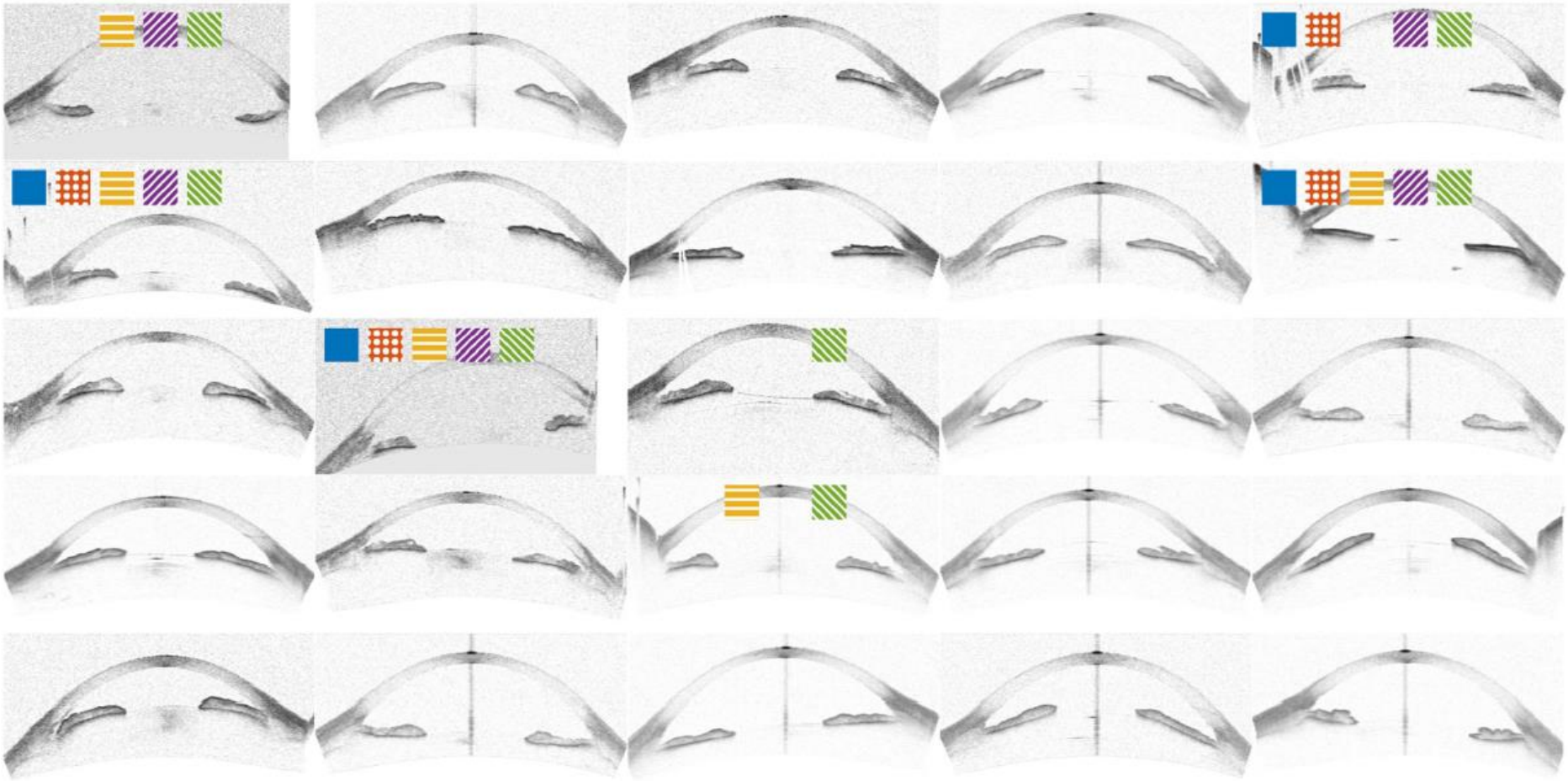
- Ramaswamy: an outlier score is assigned according to the distance of an element to its kth nearest neighbor.

- One Class Support Vector Machine (OCSVM): uses the scalar product to define a function that returns +1 in the region where normal elements are located and −1 elsewhere.

Ramaswamy et al., Efficient algorithms for mining outliers from large data sets. ACM Sigmod Record (Vol. 29, No. 2, pp. 427-438, 2000).
Schölkopf et al., Estimating the support of a high-dimensional distribution. Neural computation13, 1443-1471, 2001.

cristina.masoller@upc.edu **@cristinamasoll1**

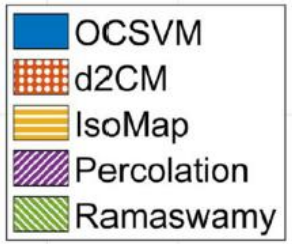
Detecting artifacts in OCT images

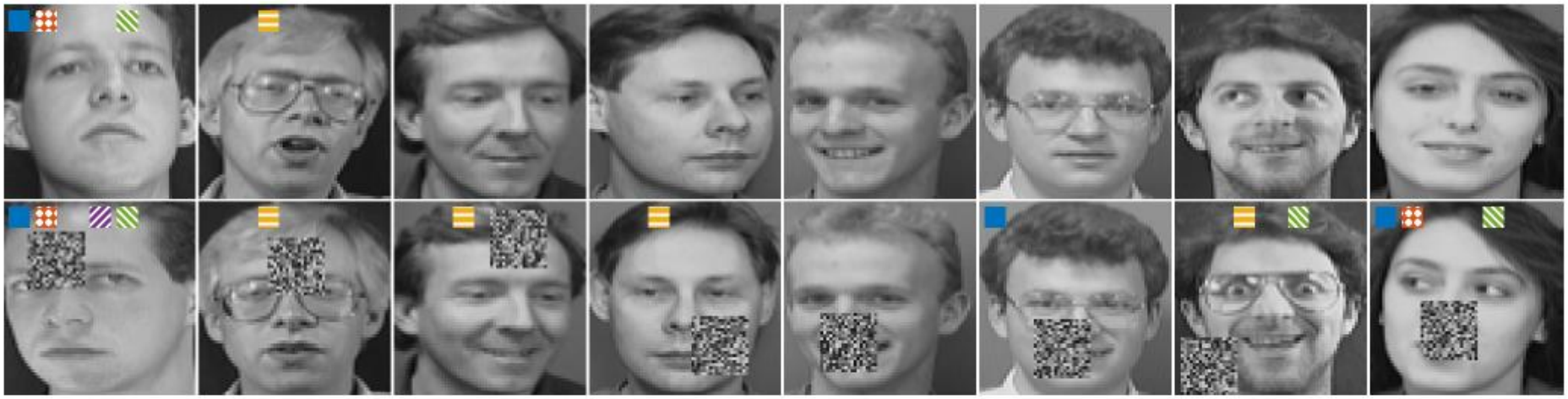


P. Amil, N. Almeida and C. Masoller, "Outlier mining methods based on graph structure analysis", Front. Phys. 7, 194 (2019).

cristina.masoller@upc.edu @cristinamasoll1

Face database





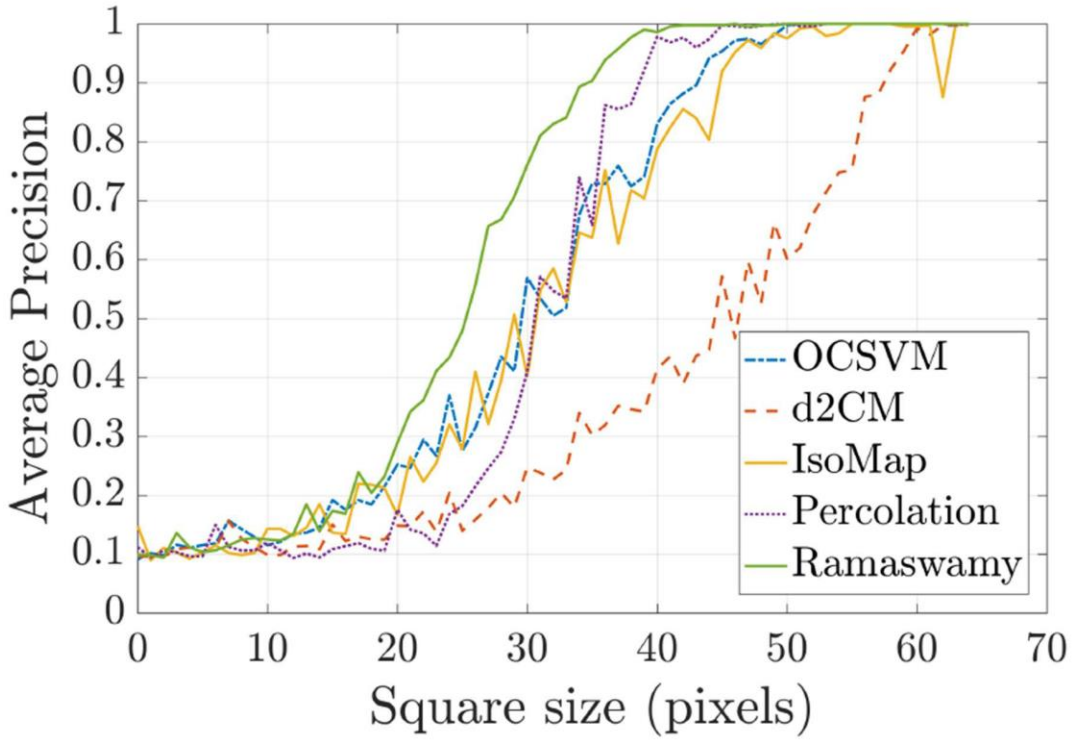
We added to some random images a square with gray-scale pixels whose color distribution is the same as that of the image.

http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html

cristina.masoller@upc.edu @cristinamasoll1

Performance quantification

Average precision: area under the Precision-Recall curve, TP/(TP+FP) vs TP

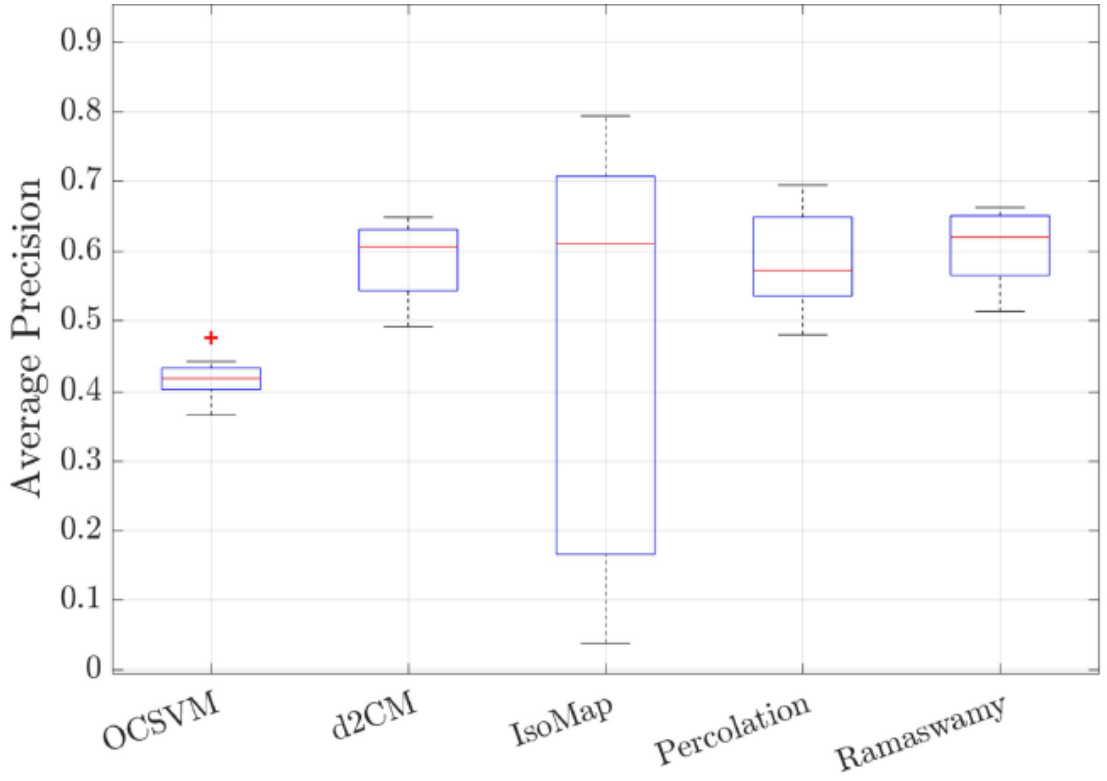


P. Amil, N. Almeida and C. Masoller, Front. Phys. 7, 194 (2019).

cristina.masoller@upc.edu @cristinamasoll1

How about other types high-dimensional of items?

We analyzed a database of Credit Card Transactions (some identified as frauds); each transaction has 28 features from PCA.



https://www.kaggle.com/mlg-ulb/creditcardfraud

P. Amil, N. Almeida and C. Masoller, Front. Phys. 7, 194 (2019).

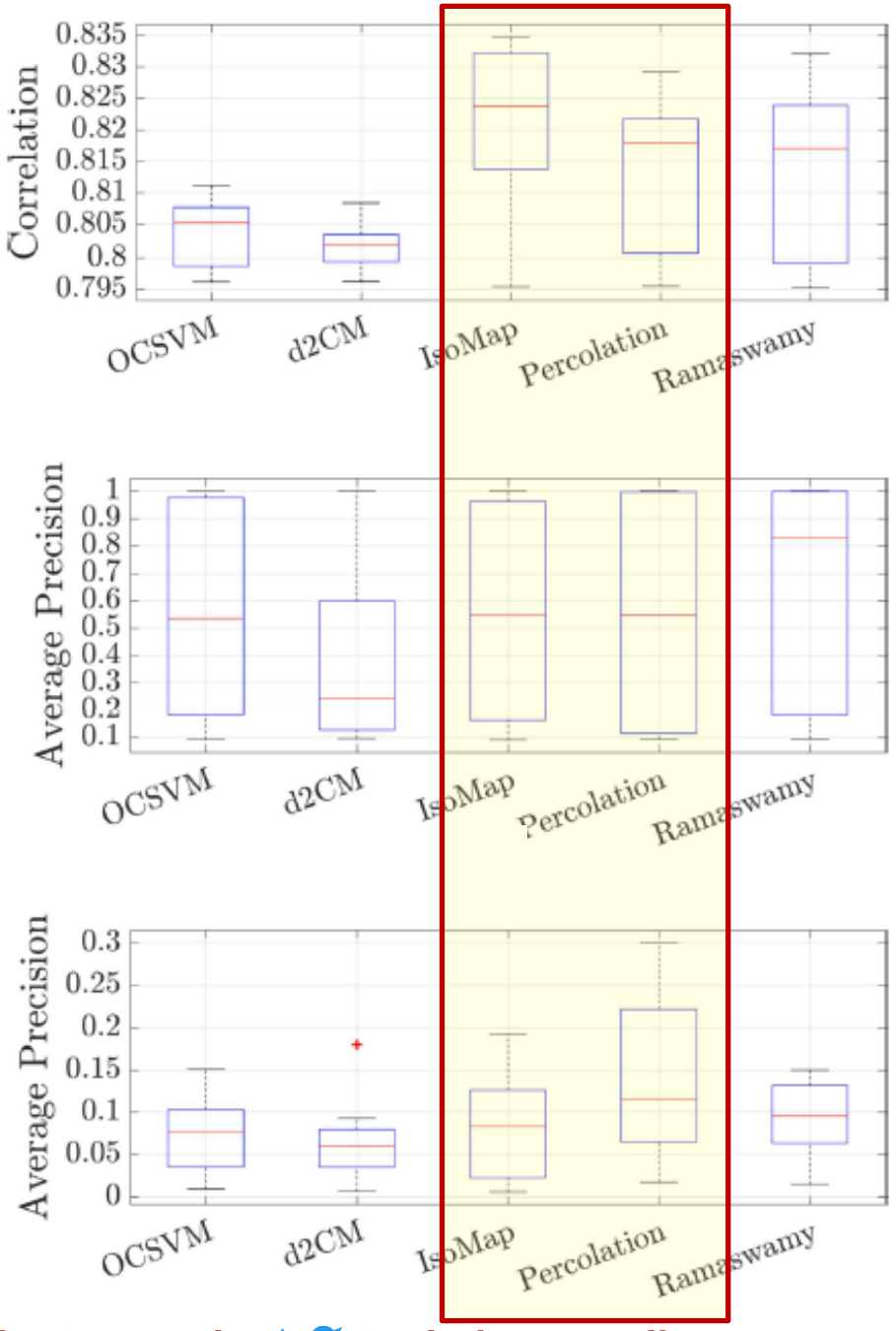
cristina.masoller@upc.edu @cristinamasoll1

Summary of results

OCT images

Face images

Credit card transactions



cristina.masoller@upc.edu @cristinamasoll1

The methods' performance depends on the data. How do they compare in terms of execution time?

For a database of 1,000 elements with 30 dimensions, run on *Matlab* on an Intel i7-7700HQ laptop:

Distance to center of mass	0.01 s
Ramaswamy	0.04 s
One Class Support Vector Machine	0.2 s
Percolation	6 s
IsoMap	18 s

Can we do better?

cristina.masoller@upc.edu @cristinamasoll1

Another distance-based way to mine anomalies / outliers

Feature vectors of items *i* and *j*

$$\{f_{i1} \dots f_{iM}\} \qquad \{f_{j1} \dots f_{jM}\}$$

1. Euclidean distance: $d_{ij} = \sqrt{\sum_{k=1}^{M} (f_{ik} - f_{jk})^2}$

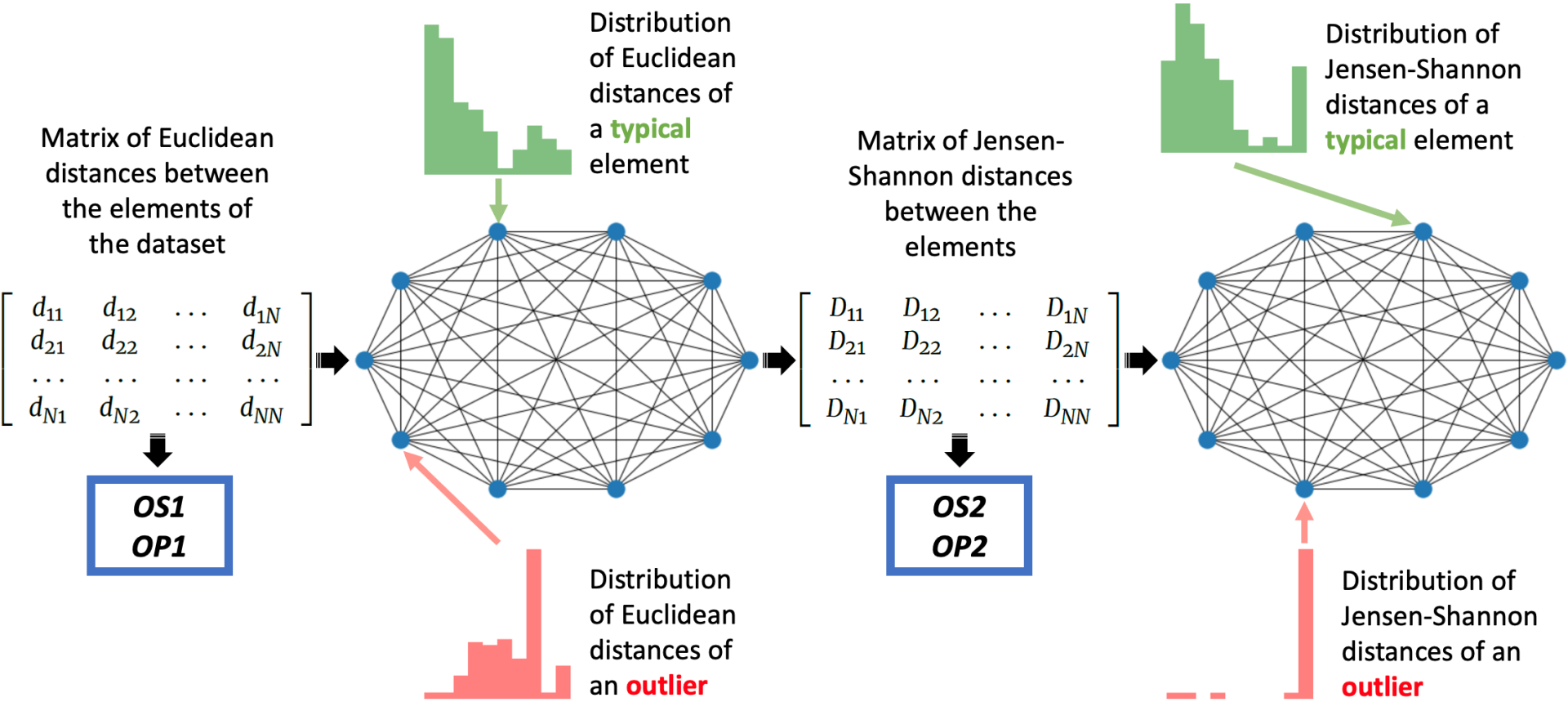
$$OS1_i = \frac{1}{N} \sum_{l=1}^{N} d_{il}$$

2. Jensen-Shannon divergence: distance of distributions of distances of images i and j: $\{d_{il}\}$ and $\{d_{jl}\}$

$$D_{ij} = JS[P_i, P_j] = H[(P_i + P_j)/2] - H[P_i]/2 - H[P_j]/2,$$

$$OS2_i = \frac{1}{N} \sum_{l=1}^{N} D_{il}$$

cristina.masoller@upc.edu @cristinamasoll1

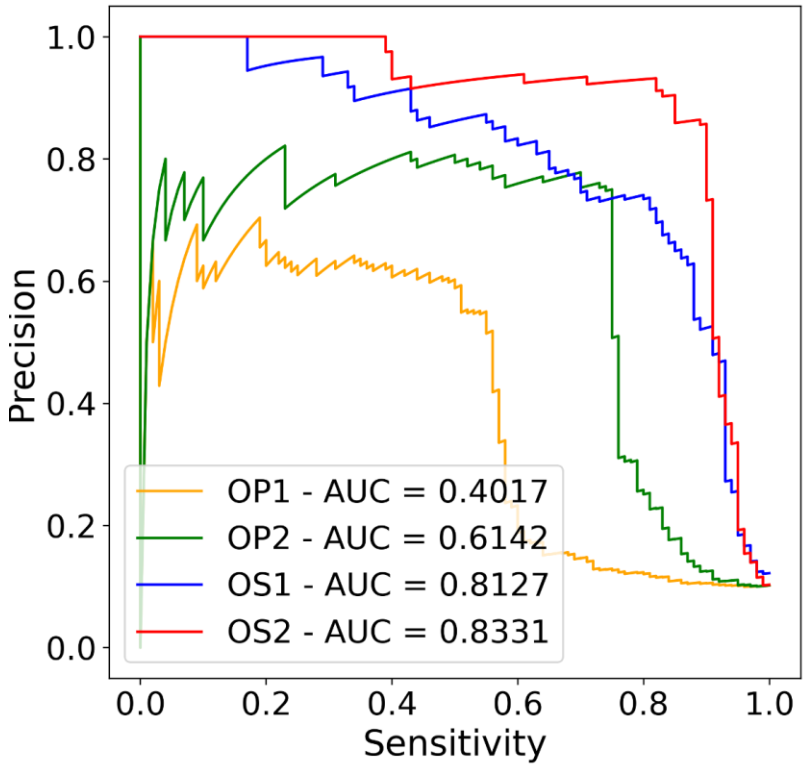


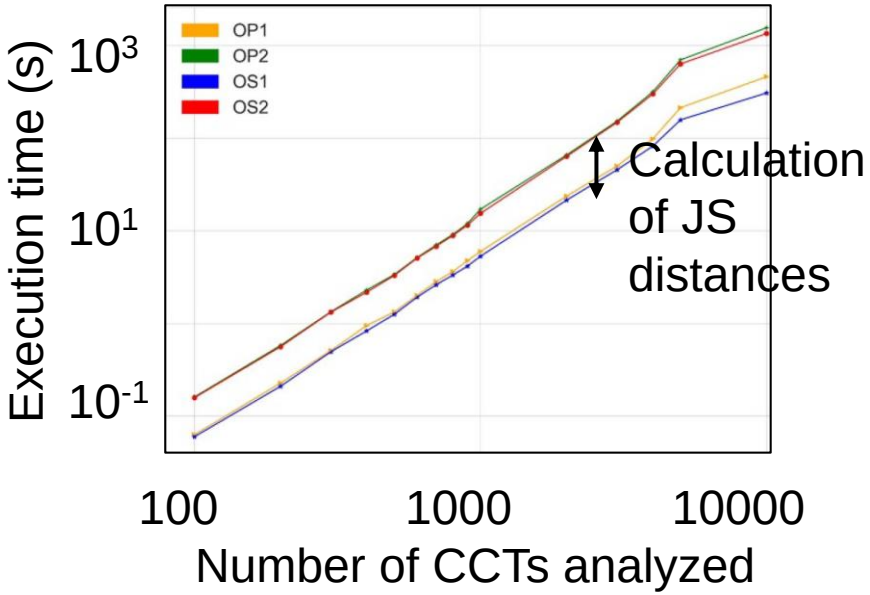
OS1, OS2: Outlier score defined by the sum of distances
OP1, OP2: Outlier score defined by the percolation order

cristina.masoller@upc.edu @cristinamasoll1

Results

For 1000 credit card transactions
10% fraud (similar with 5% fraud)





OP1: Percolation, Euclidian distances
OS1: Sum of Euclidian distances

OP2: Percolation, JS distances
OS2: Sum of JS distances

A. S. O. Toledo et. al, *Outlier mining in high-dimensional data using the Jensen-Shannon divergence and graph structure analysis*, J. of Phys: Complexity 3, 045011 (2022).

cristina.masoller@upc.edu @cristinamasoll1

Outline

- Network-based tools for retina fundus image analysis
- Anomaly / outlier detection in high dimensional data
- Ordinal (symbolic) analysis of vegetation images
- Take home messages

cristina.masoller@upc.edu @cristinamasoll1

Ordinal symbolic analysis

Bandt and Pompe: Phys. Rev. Lett. 2002

$$\{\ldots x_i, x_{i+1}, x_{i+2}, \ldots\}$$

Possible order relations among D=3 numbers (e.g., 2, 5, 7)

{…2, 5, 7…}

A 

B 

{…2, 7, 5…}

{…5, 2, 7…}

C 

D 

{…5, 7, 2…}

{…7, 2, 5 …}

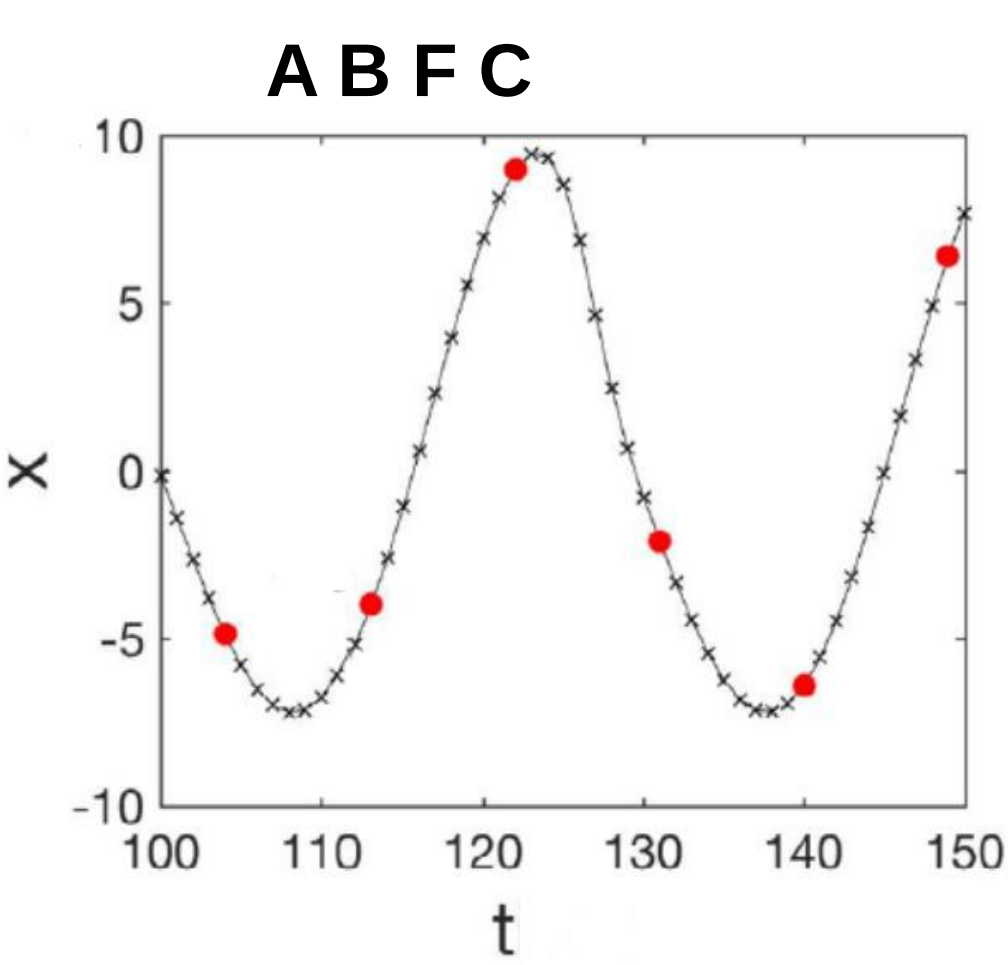
E 

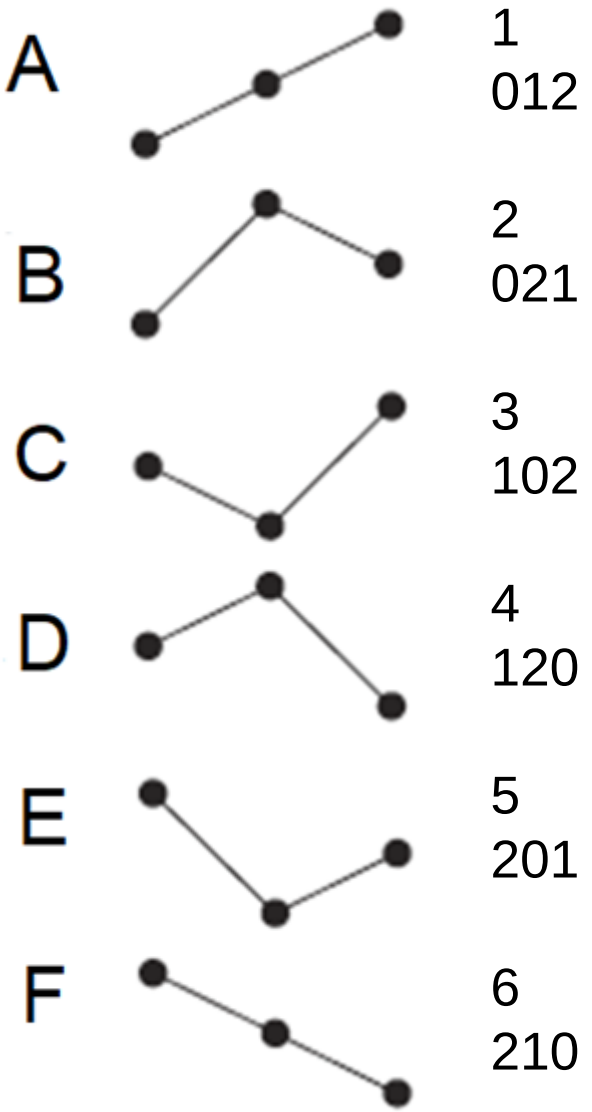
F 

{…7, 5, 2…}

cristina.masoller@upc.edu @cristinamasoll1

Using the “ordinal code”, which is the message?



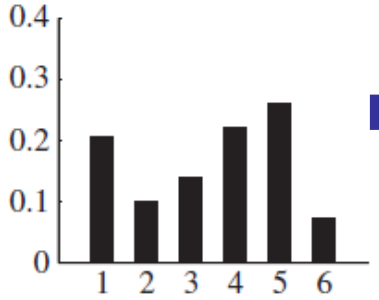


cristina.masoller@upc.edu @cristinamasoll1

Permutation entropy: Shannon's entropy computed from ordinal probabilities

Time series

Ordinal probabilities



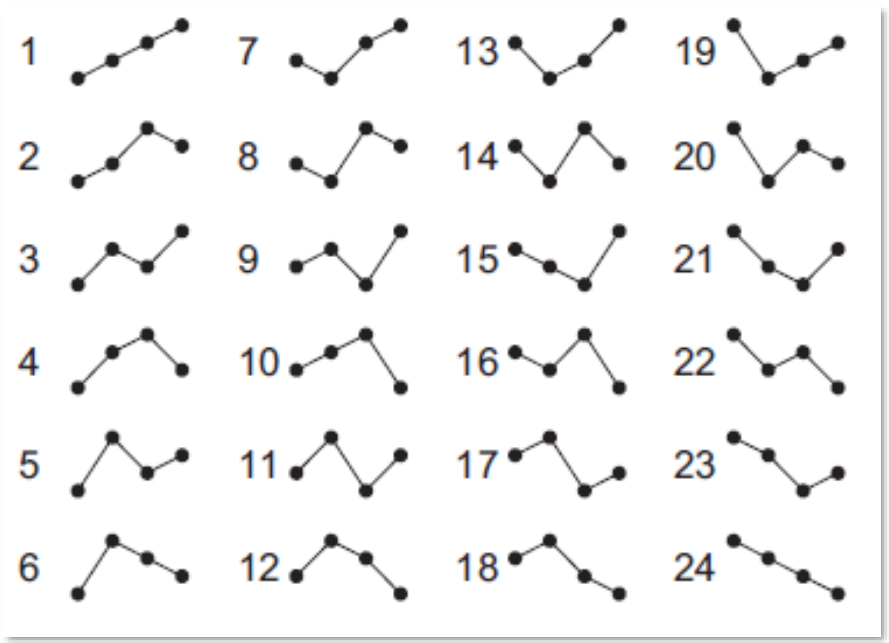
Permutation entropy

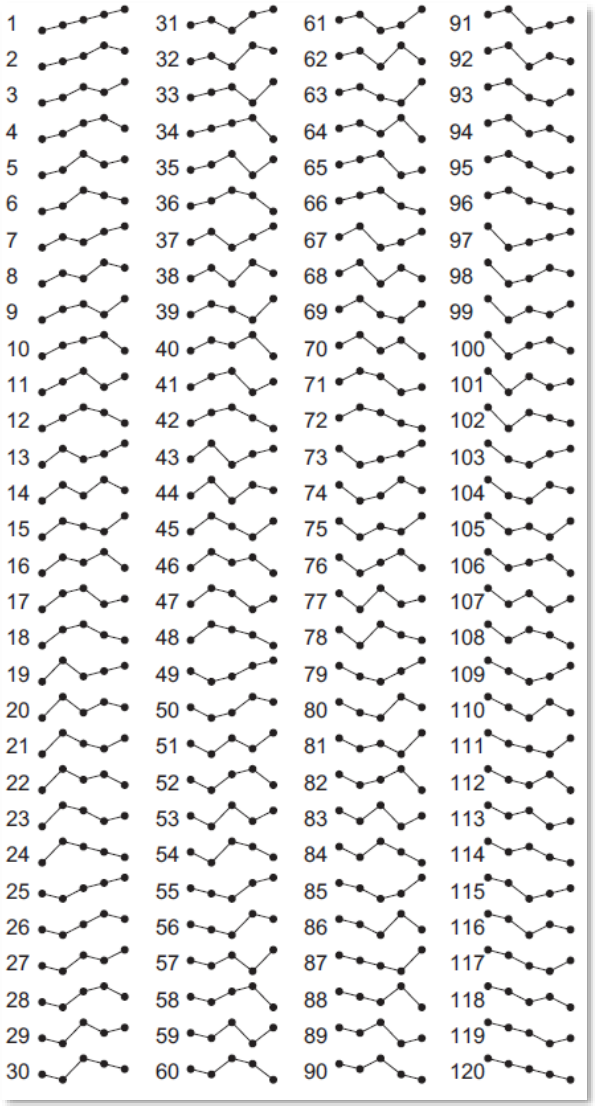
$$H = -\sum_{i=1}^{N} p_i \ln p_i$$

$$\sum_{i=1}^{N} p_i = 1$$

cristina.masoller@upc.edu @cristinamasoll1

The number of ordinal patterns increases as D!





Problem for short datasets (depending on the length of the time series, D limited to 3-5)

U. Parlitz et al. / Computers in Biology and Medicine 42 (2012) 319–327

cristina.masoller@upc.edu @cristinamasoll1

Ordinal analysis is a popular technique to analyze data generated from complex systems

- Financial, economical
- Biological, life sciences
- Geosciences, climate
- Physics, chemistry, etc.

It has been used to:

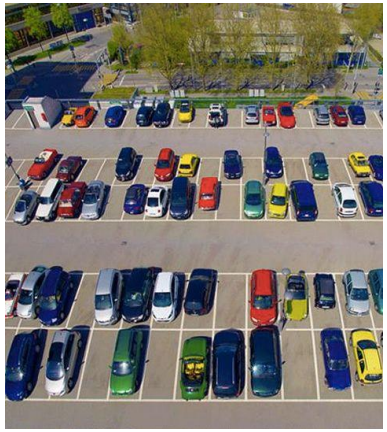
- Validate models, extract physical parameters
- Distinguish stochastic and determinism signals
- Classify different types of signals (pathological, healthy)
- Identify coupling and directionality.

I. Leyva, J. M. Martinez, C. Masoller, O. A. Rosso, M. Zanin, "*20 Years of Ordinal Patterns: Perspectives and Challenges*", EPL 138, 31001 (2022).

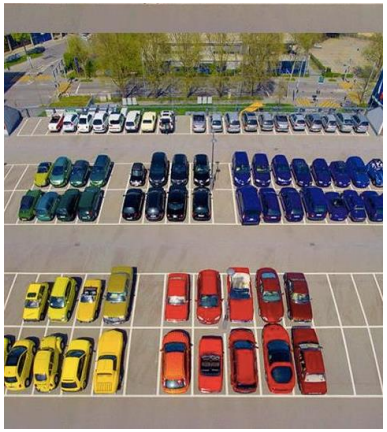
cristina.masoller@upc.edu @cristinamasoll1

Entropy, information, and complexity

High *H,* low information



Low *H,* high information



Low H -Barcelona

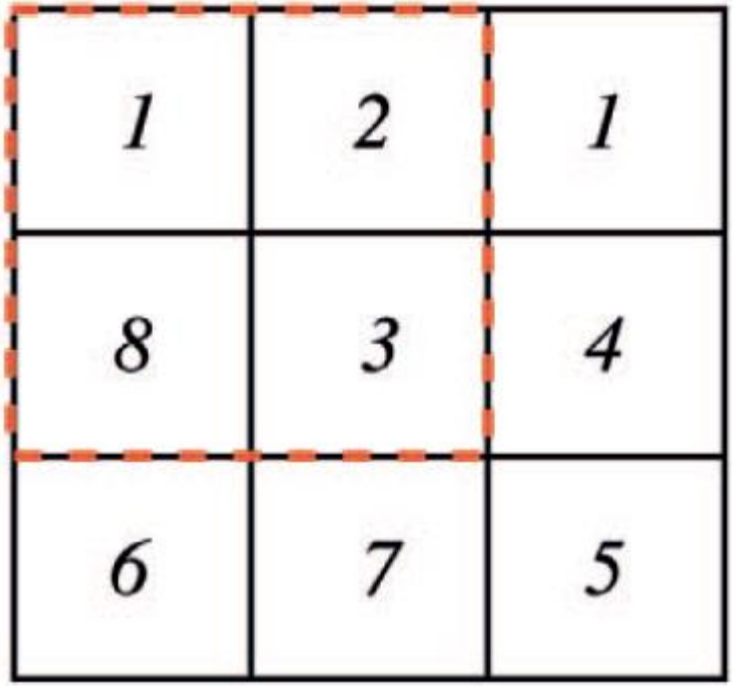
Intermediate H -Terrassa

High H –Ciudad de Mexico

J. Tiana PhD Thesis, UPC

cristina.masoller@upc.edu @cristinamasoll1

Ordinal analysis of two-dimensional patterns



1	2
8	3

1283 ⇒0132

2	1
3	4

2134 ⇒1023

8	3
6	7

8367 ⇒3012

3	4
7	5

3475 ⇒0132

Spatial permutation entropy

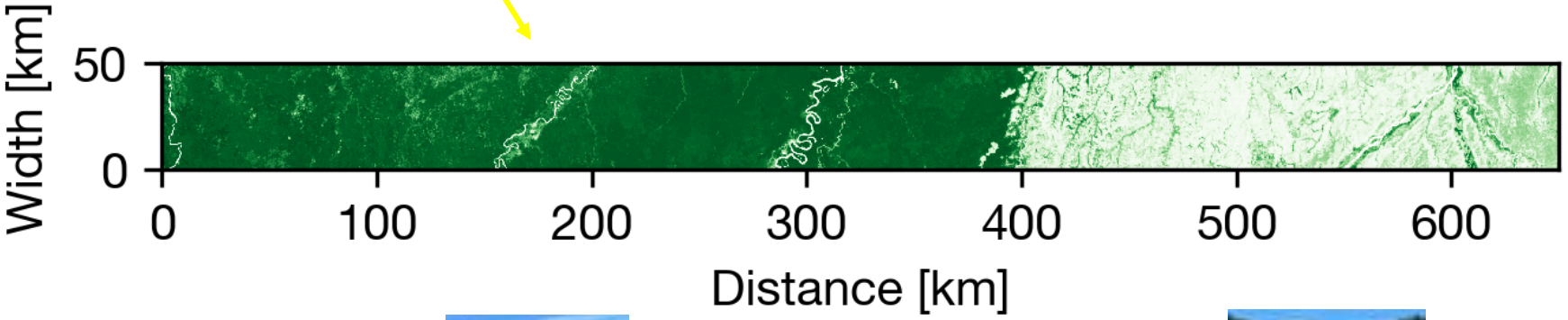
$$H = -\sum_{i=1}^{N} p_i \ln p_i$$

(2x2 pixels: 24 possible patterns)

H. V. Ribeiro et. al, PLoS ONE 7, e40689 (2012).

cristina.masoller@upc.edu @cristinamasoll1

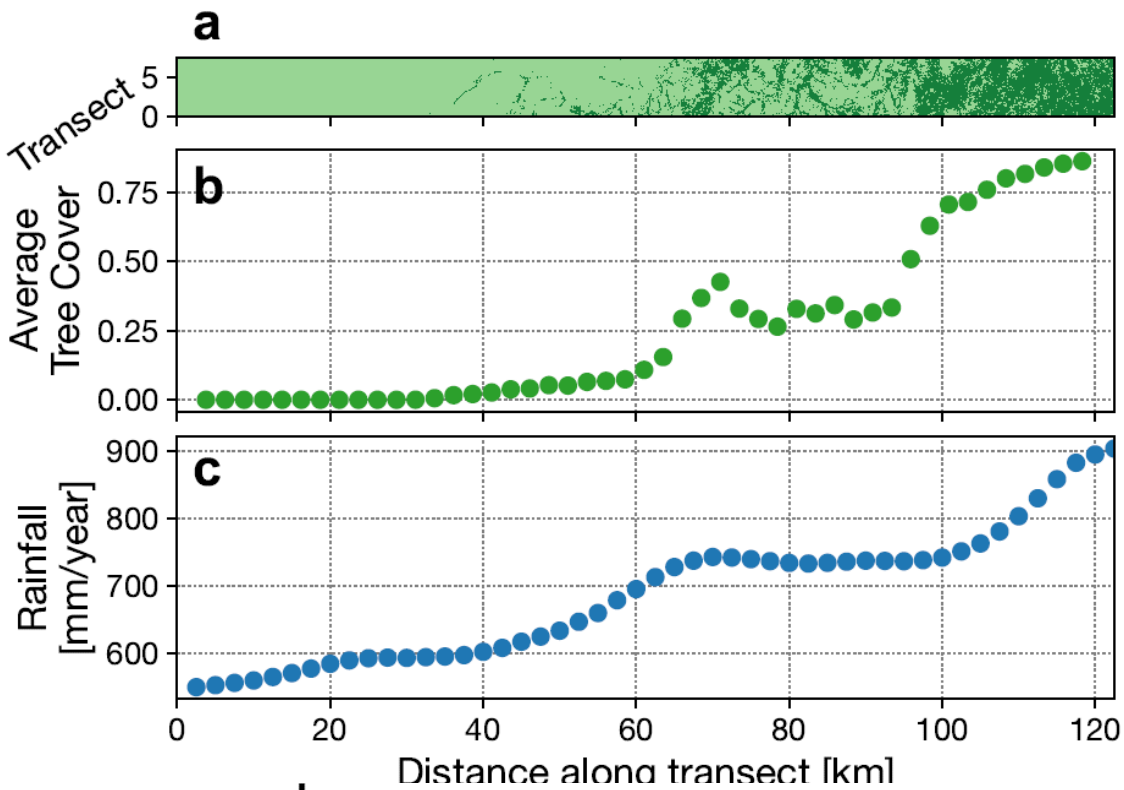
Analysis of vegetation satellite images

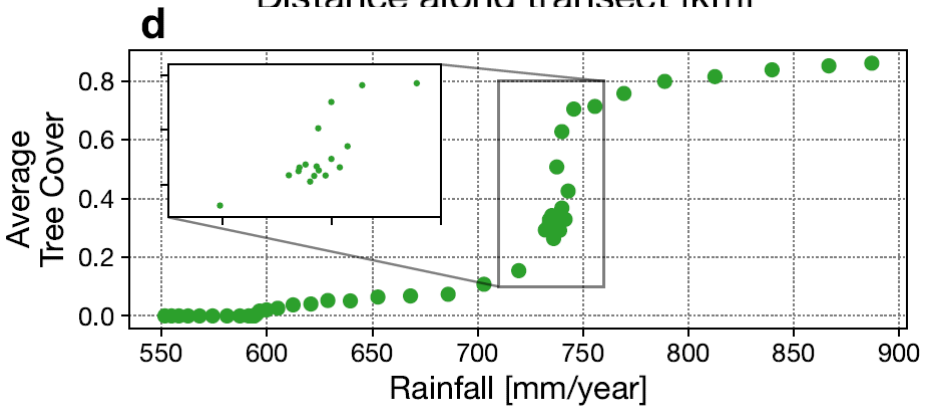


G. Tirabassi and C. Masoller, "*Entropy-based early detection of critical transitions in spatial vegetation fields*", PNAS 120, e2215667120 (2023).

cristina.masoller@upc.edu @cristinamasoll1

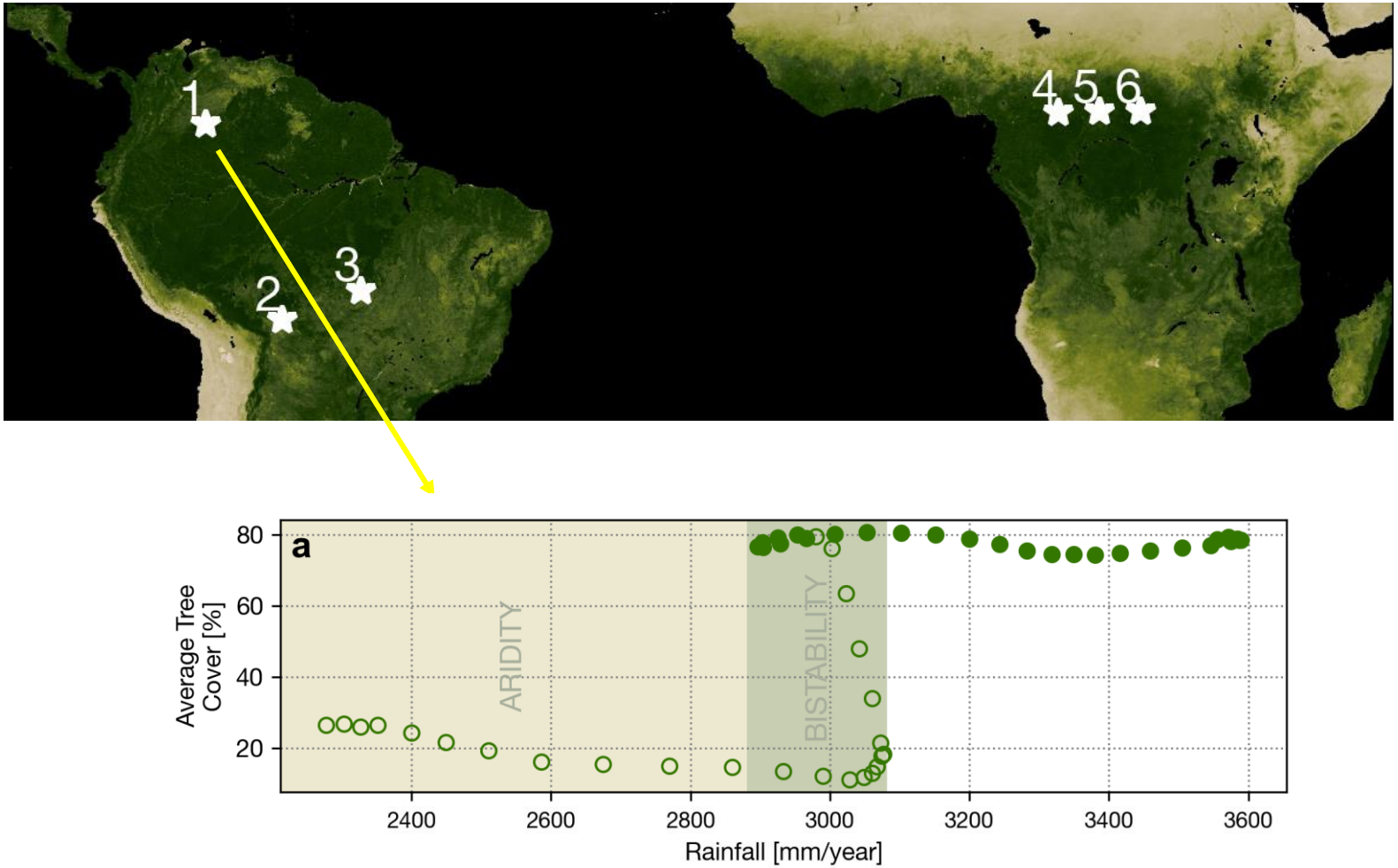
Spatial tree cover depends on the annual rainfall





cristina.masoller@upc.edu @cristinamasoll1

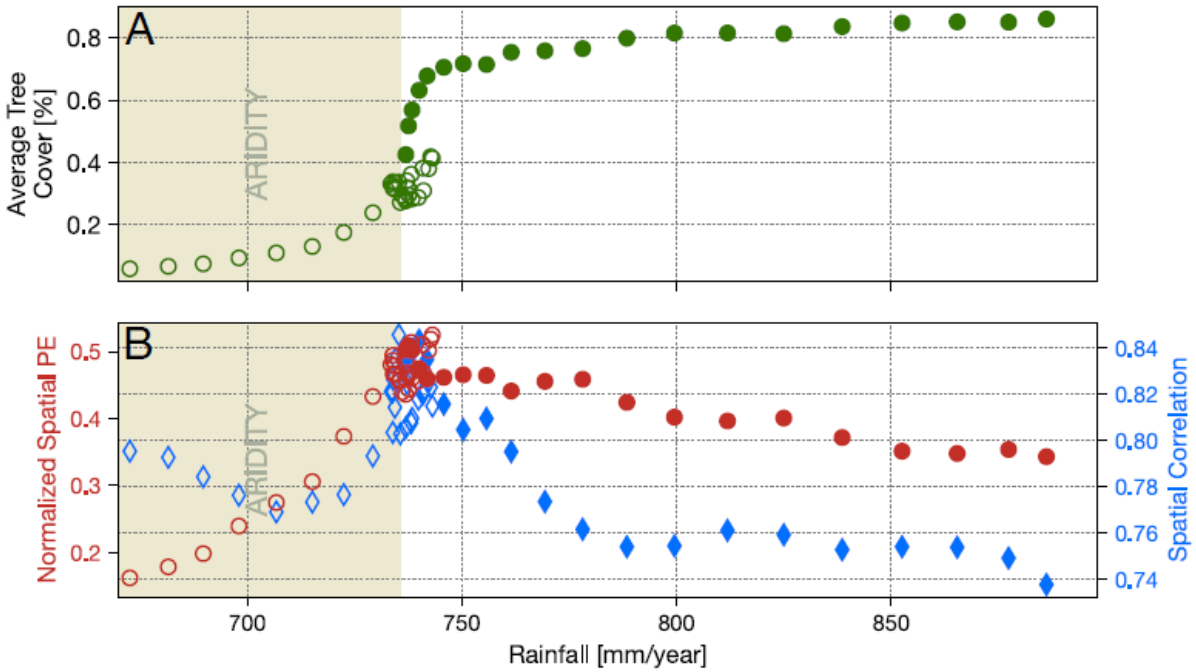
Can we try to anticipate the desertification transition?



G. Tirabassi and C. Masoller, "*Entropy-based early detection of critical transitions in spatial vegetation fields*", PNAS 120, e2215667120 (2023).

cristina.masoller@upc.edu @cristinamasoll1

The variation of the spatial permutation entropy can provide an indication of the approaching vegetation transition



Spatial correlation: Moran's I coefficient

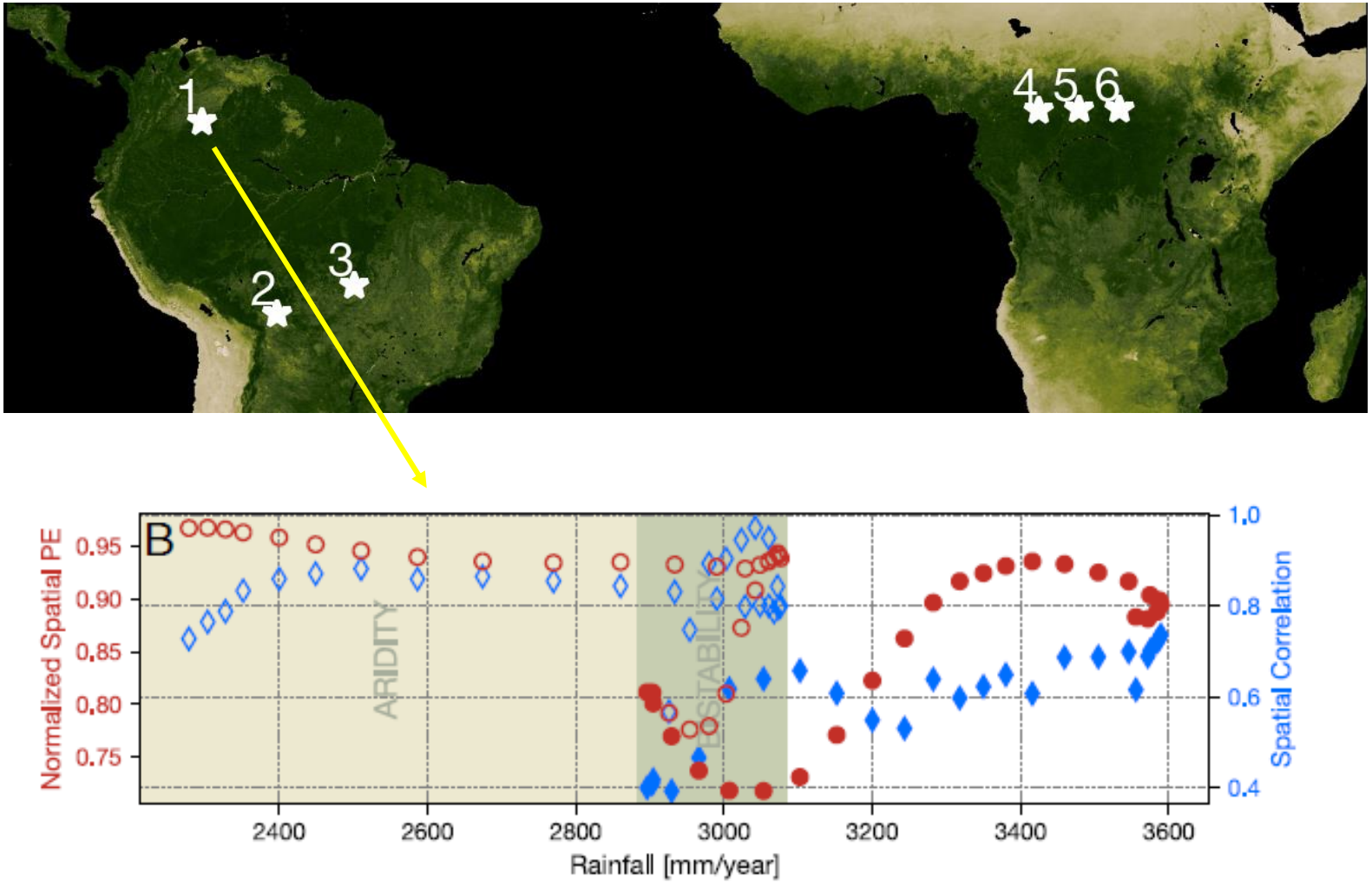
$$I = \frac{N}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij}(u_i - \bar{u})(u_j - \bar{u})}{\sum_i (u_i - \bar{u})^2}, \qquad [2]$$

where the coefficients w_{ij} are 1 if the points i and j are first neighbors and 0 otherwise, $\bar{u}$ is the average value of u_{ij}, and N is the total number of points.

G. Tirabassi and C. Masoller, "*Entropy-based early detection of critical transitions in spatial vegetation fields*", PNAS 120, e2215667120 (2023).

cristina.masoller@upc.edu @cristinamasoll1

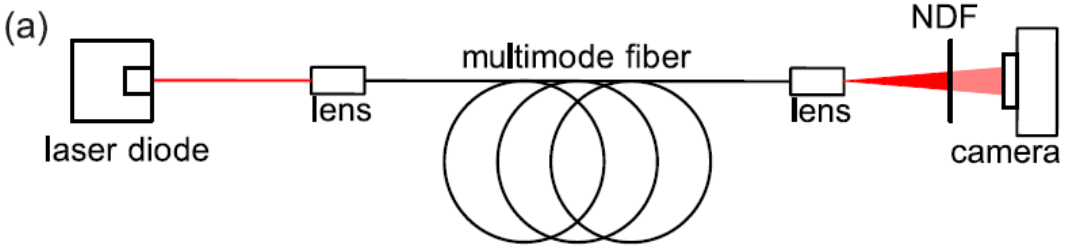
Can we try to anticipate the desertification transition?

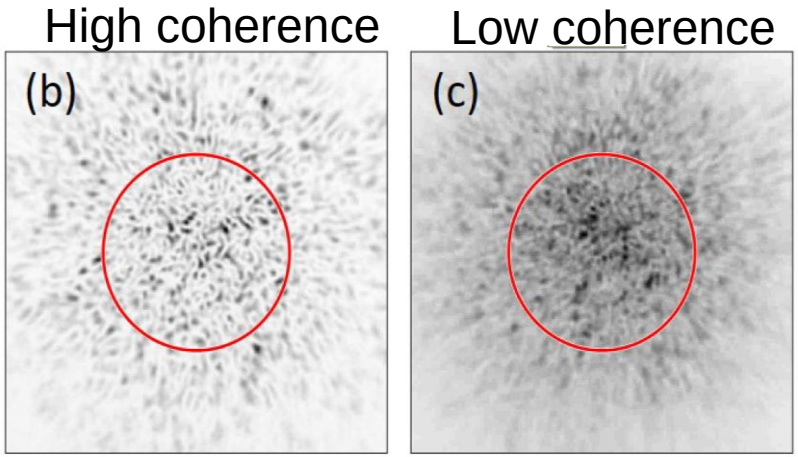


G. Tirabassi and C. Masoller, "*Entropy-based early detection of critical transitions in spatial vegetation fields*", PNAS 120, e2215667120 (2023).

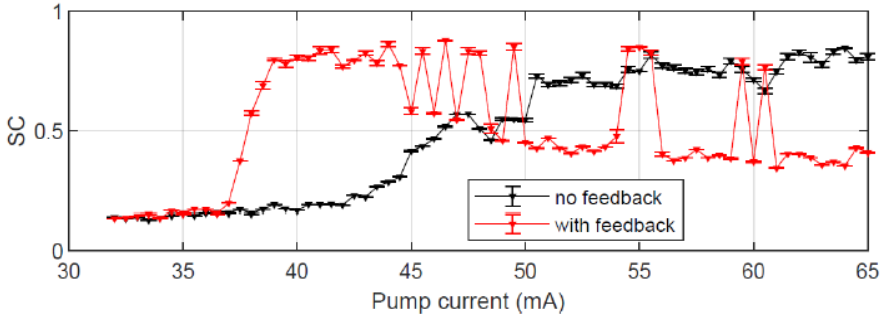
cristina.masoller@upc.edu @cristinamasoll1

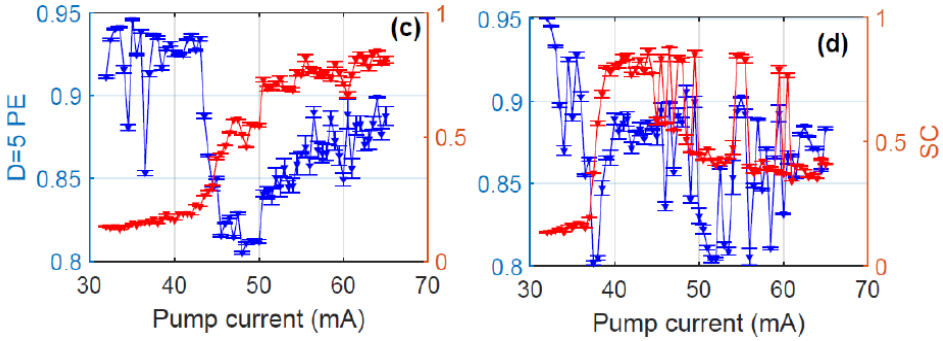
Extra bonus: can spatial permutation entropy anticipate the turn on of a laser?





$$SC = \sigma / \langle I \rangle$$





G. Tirabassi et al, "*Permutation entropy-based characterization of speckle patterns generated by semiconductor laser light*", Submitted (2023)

cristina.masoller@upc.edu @cristinamasoll1

Take home messages

- Retina fundus image analysis: network-based methods return informative features that may allow an early identification of ophthalmic diseases.
- Outlier mining: “Distance-based” methods identify outliers/artifacts in high-dimensional, not-too-large databases (images, credit card frauds). The execution time grows linearly with the dimension of the data (# of features) and at least as NxN with the size of the database.
- Spatial permutation entropy computed from images may return useful information to identify transitions that occur under changing conditions or varying parameters.
- The performance of all these methods depends on the characteristics of the data.

cristina.masoller@upc.edu @cristinamasoll1

Thanks to

Pablo Amil, Alex S. O. Toledo, Giulio Tirabassi

References:

- P. Amil et al., "*Network-based methods for retinal fundus image analysis and classification*", PLoS ONE 14 e0220132 (2019).
- P. Amil et al., "*Outlier mining methods based on network structure analysis*", Front. Phys. 7, 194 (2019).
- A. S. O. Toledo et. al, "*Outlier mining in high-dimensional data using the Jensen-Shannon divergence and graph structure analysis"*, J. of Phys: Complexity 3, 045011 (2022).
- G. Tirabassi, C. Masoller, "*Entropy-based early detection of critical transitions in spatial vegetation fields*", PNAS 120, e2215667120 (2023).

Thank you for your attention !

cristina.masoller@upc.edu @cristinamasoll1