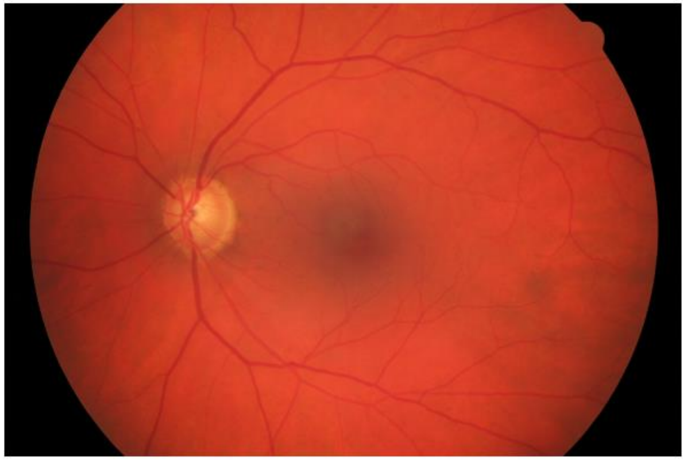
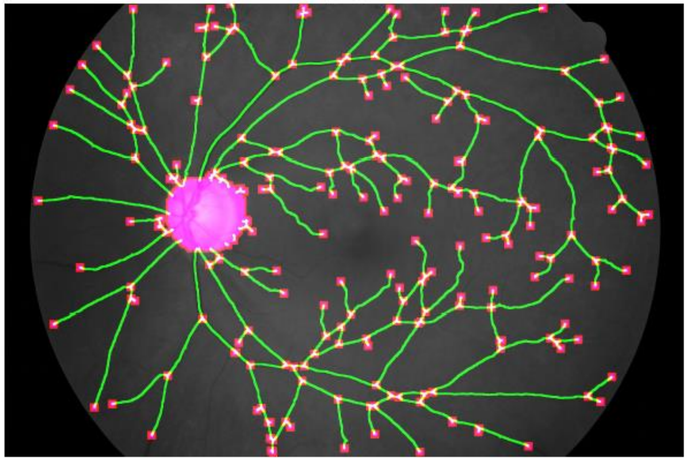




Application of network techniques to ophthalmic image analysis and outlier detection

Pablo Amil, Cristina Masoller

Universitat Politècnica de Catalunya

Terrassa, Barcelona, Spain

www.fisica.edu.uy/~cris

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

Campus d'Excel·lència Internacional

11th International Conference on Network Analysis (NET 2021)

Nizhny Novgorod, October 18, 2021

GOBIERNO DE ESPAÑA MINISTERIO DE ECONOMÍA Y COMPETITIVIDAD

ICREA

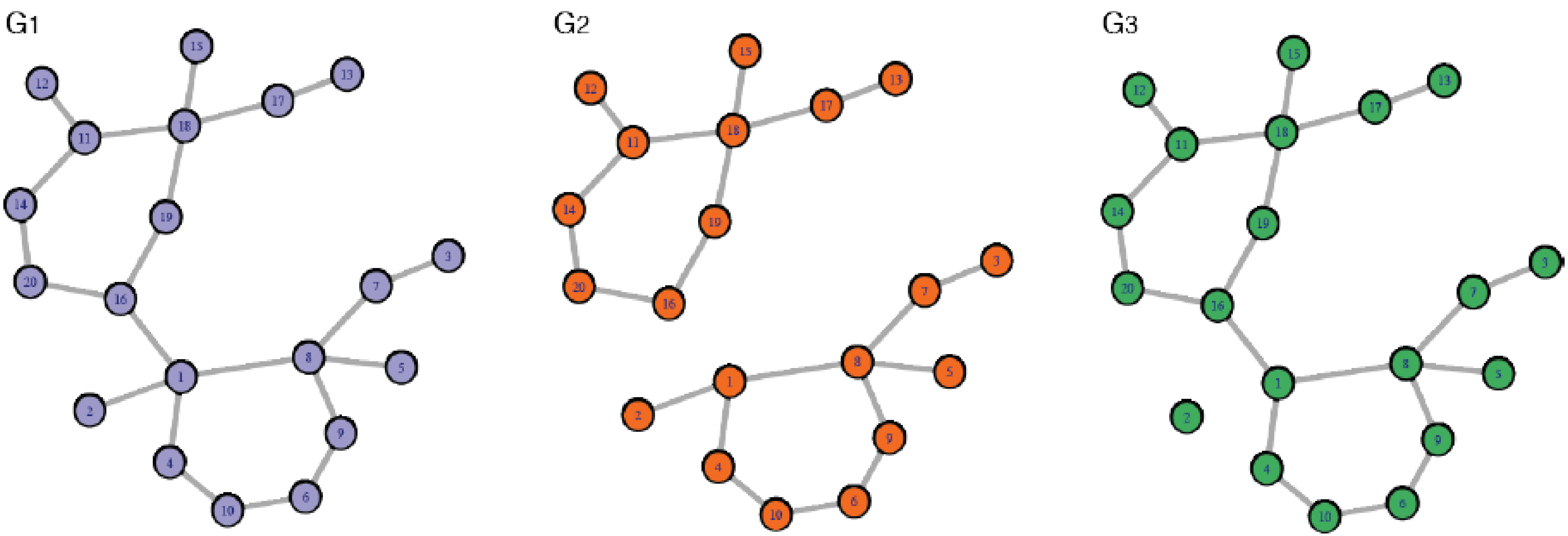
Outline

- How to detect and to quantify structural differences between networks?
  - Application to the analysis of EEG signals
  - Application to ophthalmic images
- How to identify outliers in high-dimensional data?
  - Application to ophthalmic images
  - Other applications

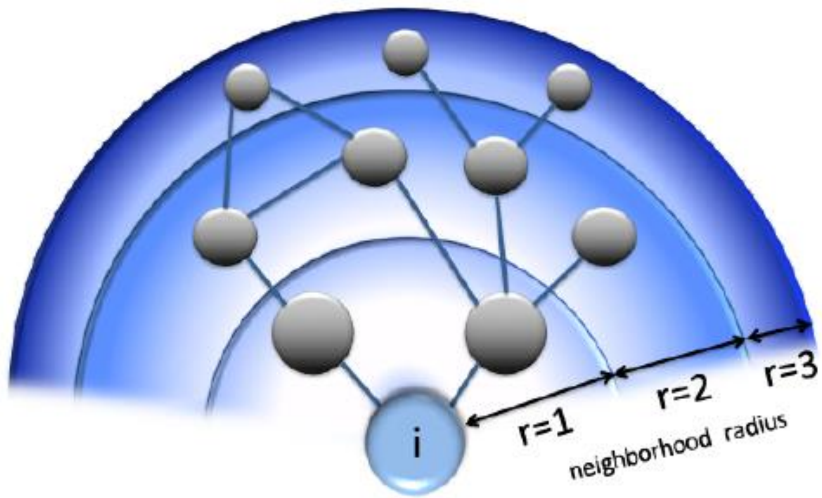
To detect structural differences between networks we need a precise measure to compare them.

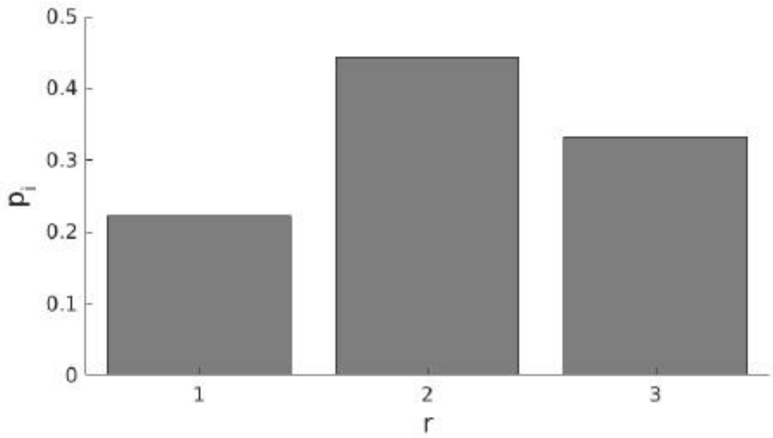
- Degree, centrality, assortativity distributions etc. provide *partial* information.

- Main problem: not all the links have the same importance.



Node Distance Distribution (NDD) of node i: fraction of nodes that are connected (shortest path) to i at distance j.





- A network with N nodes:

  NDDs = vector with N distributions

- If two networks have the same NDDs $\Rightarrow$ they have the same diameter, average path length, etc.

Adapted from A. Viol et al. Entropy 2019

How to summarize the information contained in the node distance distributions?

The *Network Node Dispersion (NND*) quantifies the heterogeneity of the distributions $\{p_1, p_2, \ldots, p_N\}$

average NDD: $\mu = \langle p_i \rangle_i$

Kullback distance between p_i and μ: $J(p_i, \mu) = \sum_r p_i(r) \log\left(\frac{p_i(r)}{\mu(r)}\right)$

$$NND(G) = \frac{\langle J(p_i, \mu) \rangle_i}{\log(d+1)}$$

d = diameter

Dissimilarity measure

$$D(G,G') = w_1\sqrt{\frac{\mathcal{J}(\mu_G,\mu_{G'})}{\log 2}} + w_2\left|\sqrt{\mathrm{NND}(G)} - \sqrt{\mathrm{NND}(G')}\right|$$

$w_1=w_2=0.5$

compares the averaged connectivity

compares the heterogeneity of the connectivity distances

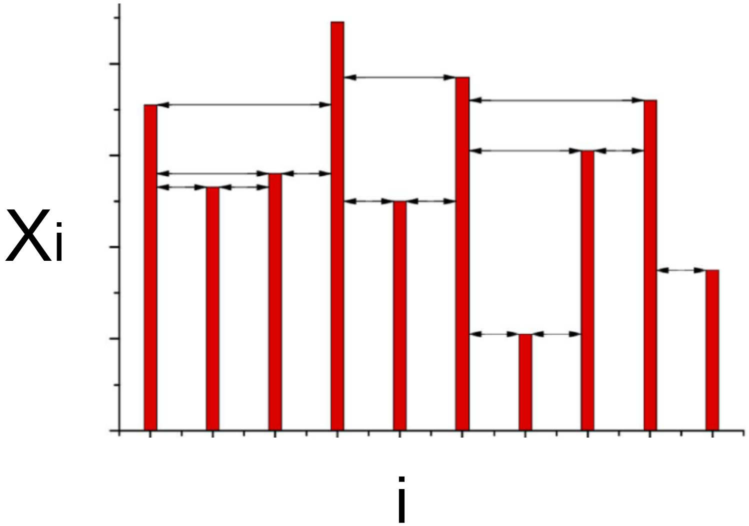
- Extensive numerical experiments demonstrate that isomorphic graphs return ***D=0.***
- Computationally efficient.
- Can be used to compare graphs with different number of nodes.

T. A. Schieber, L. Carpi, A. Diaz-Guilera, P. M. Pardalos, C. Masoller, M. G. Ravetti, "*Quantification of network structural dissimilarities*", Nat. Comm. 8, 13928 (2017).

First example of application: classification of EEG signals

Data and methodology

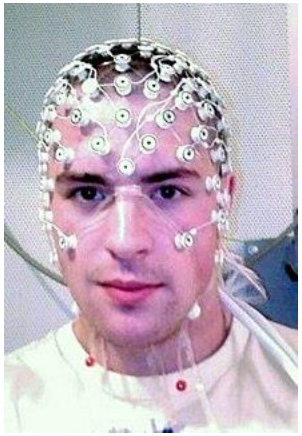
- EEG data (*)
  - 64 electrodes placed on the subject's scalp sampled at 256 Hz during 1s
  - 107 subjects: 39 control and 68 alcoholic
- Each EEG TS is transformed into a graph G using the horizontal visibility rule.

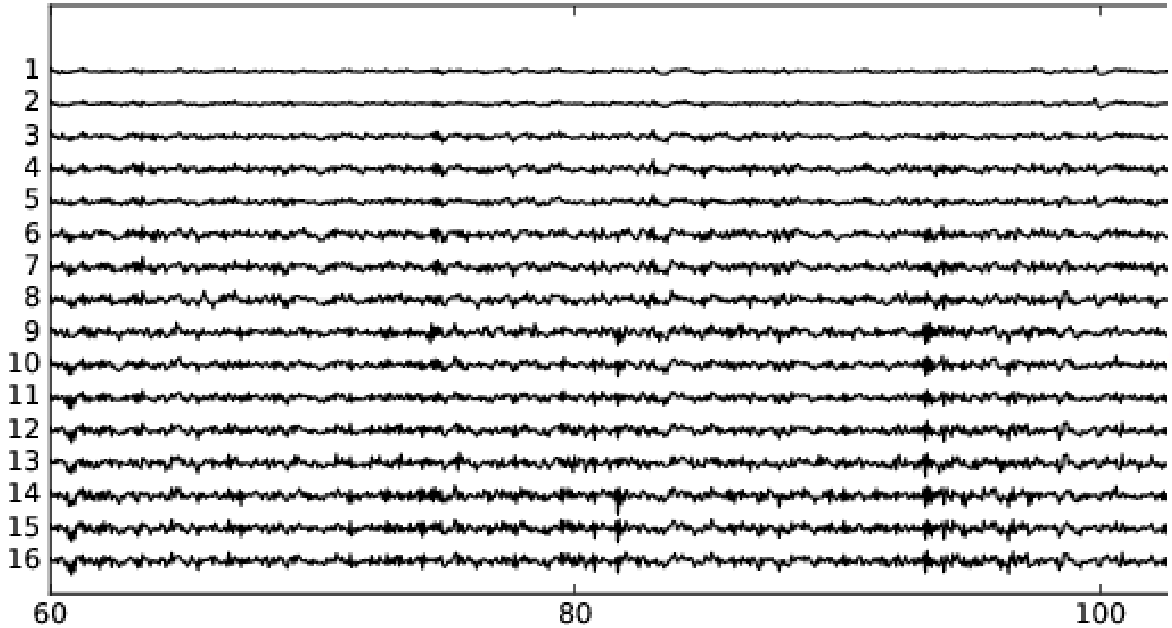


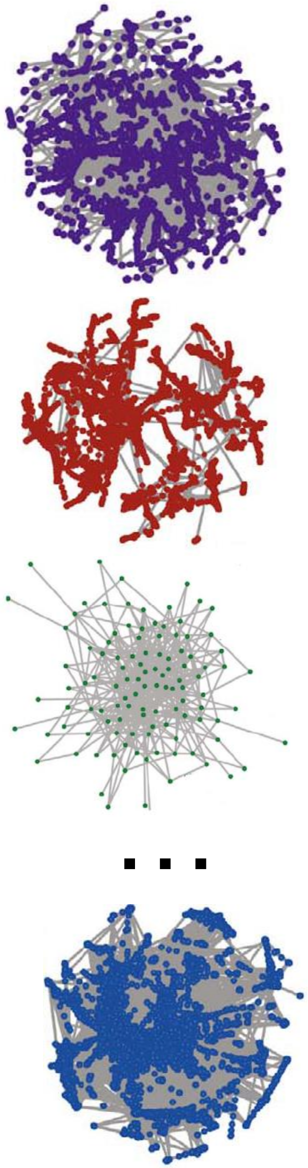
- Data points i and j are connected if there is "visibility" between them.
- We obtain an unweighted and undirected graph.
- Parameter free!

* *https://archive.ics.uci.edu/ml/datasets/eeg+database*
HVG method: Luque et al PRE (2009); Gomez Ravetti et al, PLoS ONE (2014)

For each subject dataset has 64 channels ⇒ 64 networks

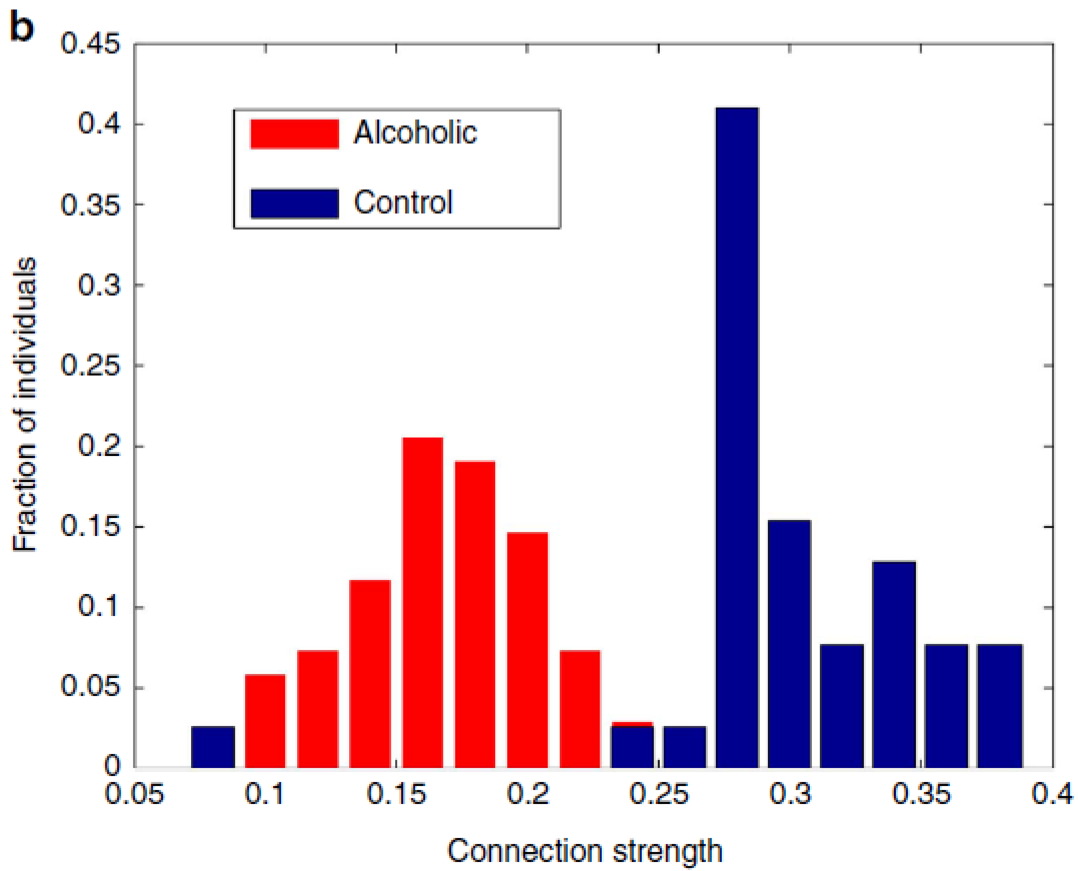






- The weight of the link between two brain regions is: **1-*D*(G,G')**

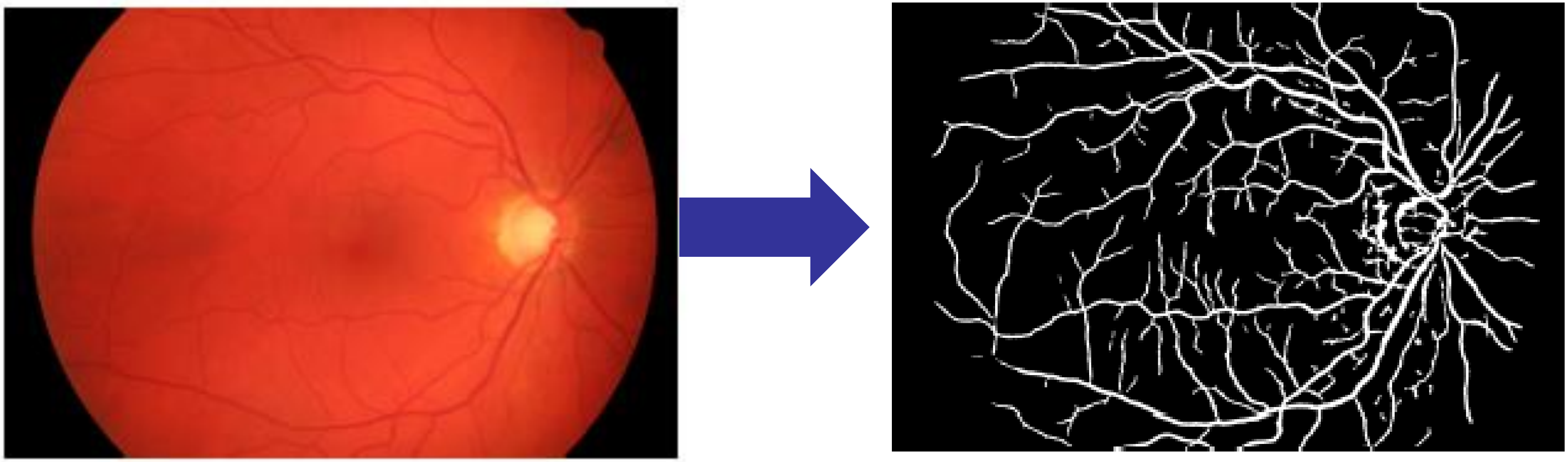
We identify two brain regions (called ‘nd’ and ‘y’), where the connection strength between these regions is higher in control than in alcoholic subjects.



T. A. Schieber, L. Carpi, A. Diaz-Guilera, P. M. Pardalos, C. Masoller, M. G. Ravetti, “*Quantification of network structural dissimilarities*”, Nat. Comm. 8, 13928 (2017).

Second application:
classification of retina images

Segmentation

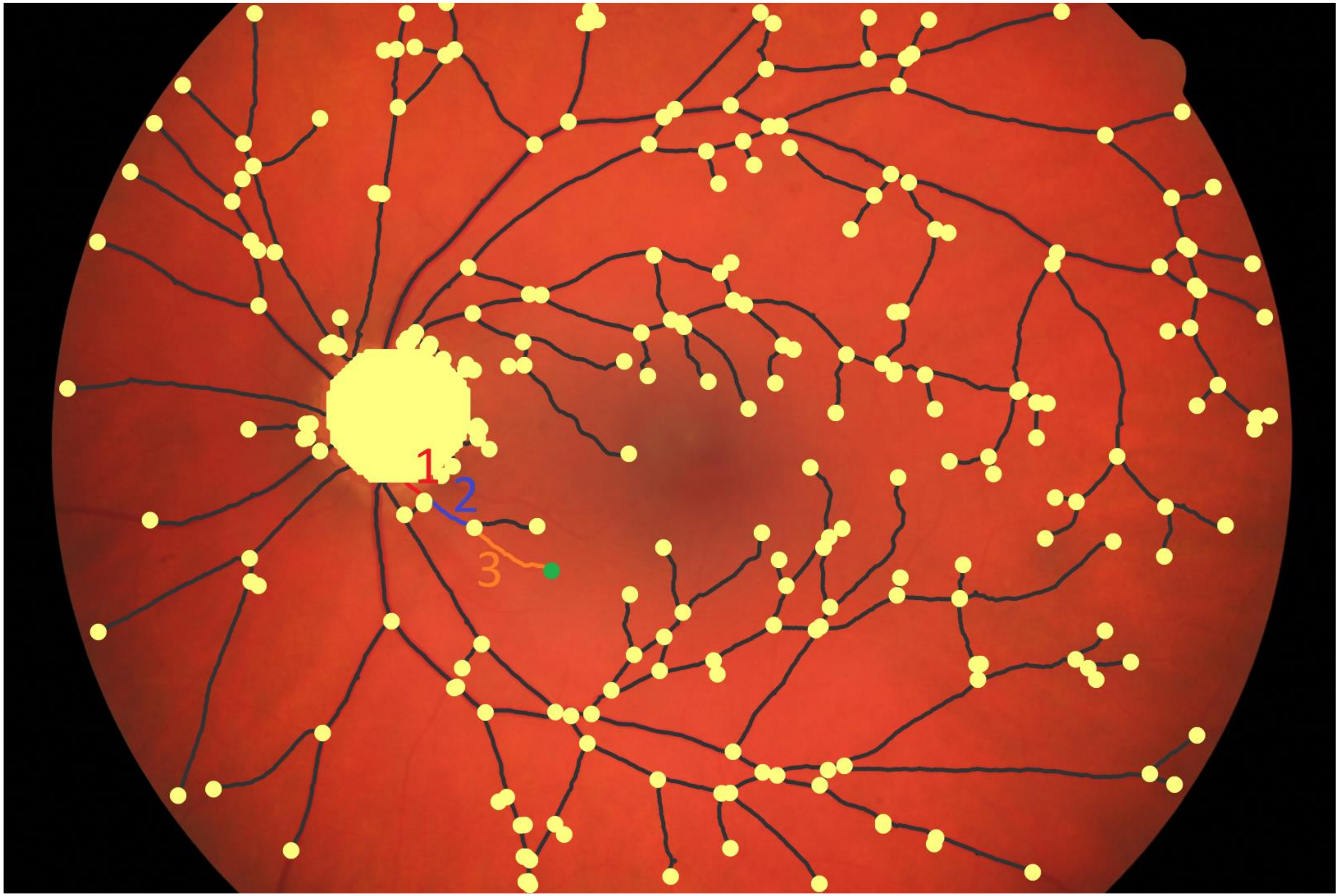


Unsupervised segmentation algorithm adapted from an algorithm used from the analysis of cultured neuronal networks.

Problem: arteries and veins networks collapse.

Santos-Sierra D, Sendiña-Nadal I, Leyva I, et al. *Graph-based unsupervised segmentation algorithm for cultured neuronal networks' structure characterization and modeling*. Cytometry Part A. 87, 513 (2015).

Identification of nodes and links.
Then, analysis of the connectivity paths to the central node (optical nerve)

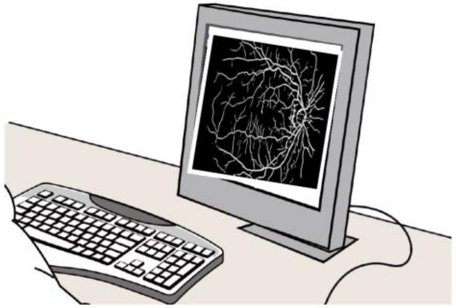


Data (1/2)

- High-resolution public database with
  - 15 healthy subjects
  - 15 glaucoma
  - 15 diabetic retinopathy

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

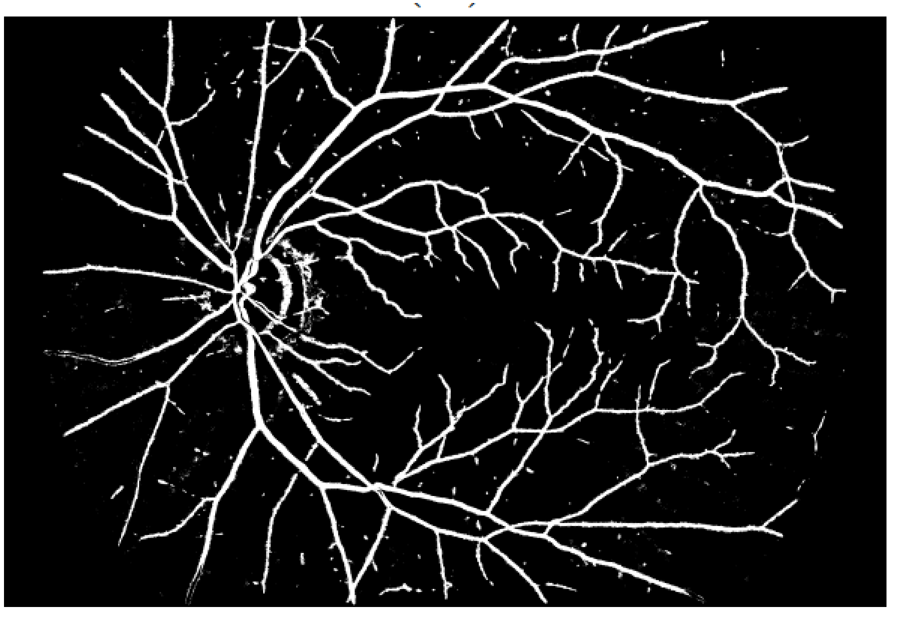
- From each fundus photography $\Rightarrow$ automated segmentation.



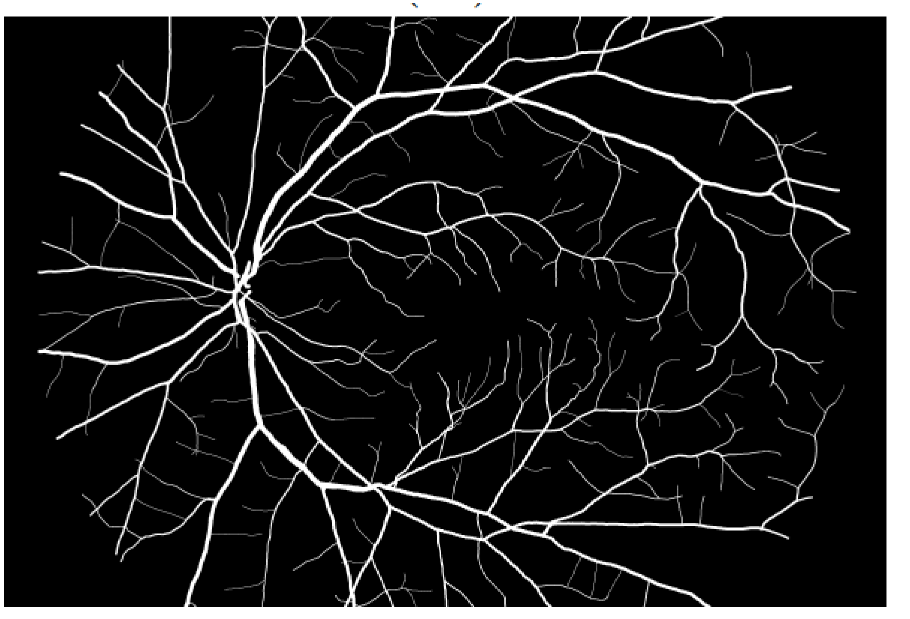
https://www5.cs.fau.de/research/data/fundus-images/

Comparison

Automated



Manual



Data (2/2)

- For both segmentations (manual and automated) we analyzed
  - raw segmented image (i.e., a binary image that includes all the pixels that correspond to vessels);
  - skeletonized image (i.e., a binary image where the width of each vessel segment is reduced to one pixel, without changing the length, location and orientation of the segment).
- We also analyzed two larger databases but with lower resolution. Best results obtained for high-resolution images.

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 perhaps affects 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).

Methodology (1/2)

From each image we calculate

- Fractal dimension (raw and skeletonized segmented images)
- Distribution of distances to the central node
- Distribution of average weights along the path to the central node
- Distribution of weighted degree

Methodology (2/2)

- We use the Jensen-Shannon (JS) divergence to compare distributions (image “i” with all other images)

- For each image “i” we obtain a feature vector

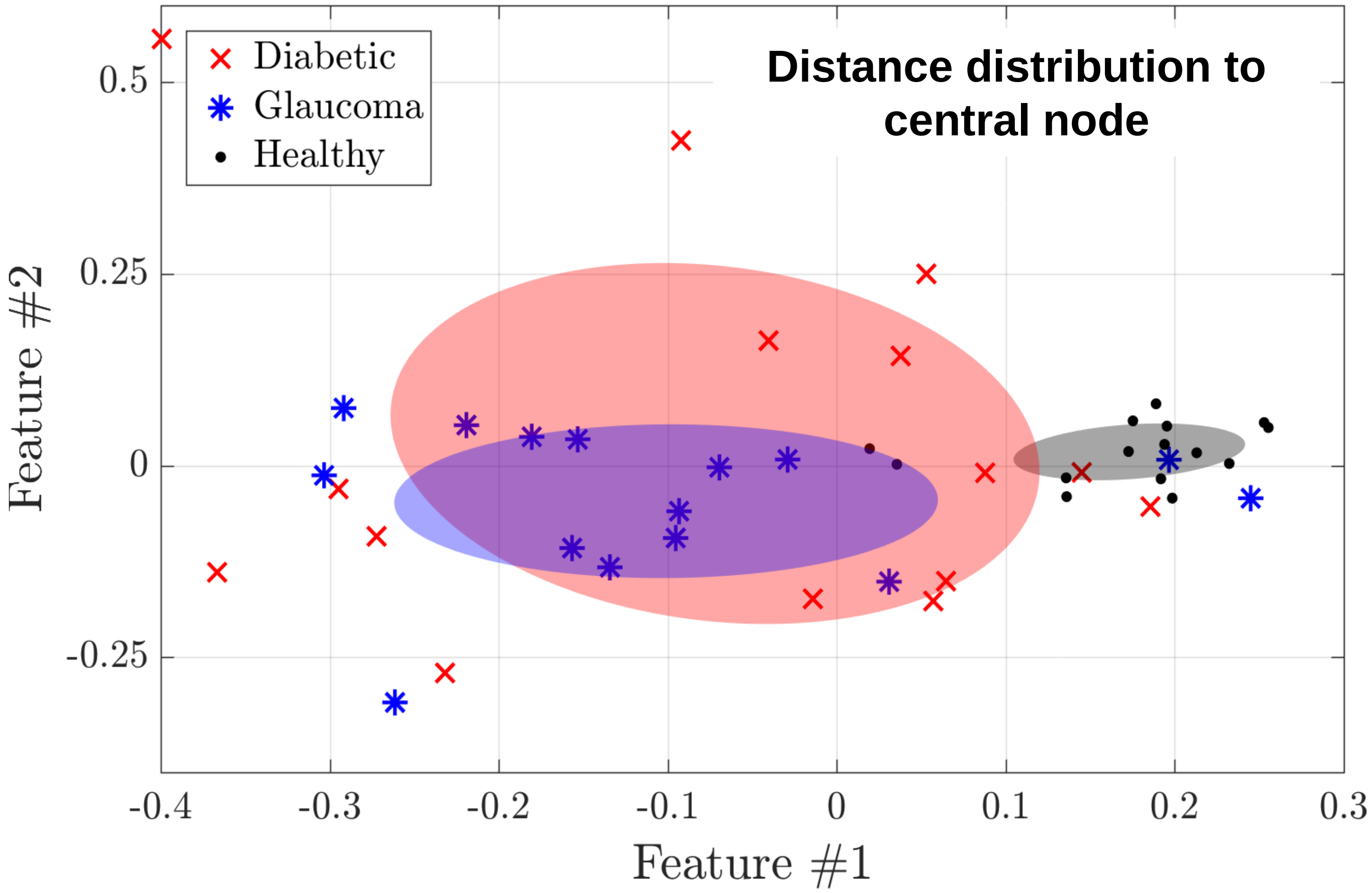
$$\{d_{i1}, d_{i2}, \ldots d_{iN}\} \qquad \text{(N = number of images)}$$

  whose elements are the distances between the distributions extracted from image “i” and image “j” (j in 1…N).

- We apply a nonlinear dimensionality reduction algorithm (*IsoMap*) to obtain only 2 features for each image.

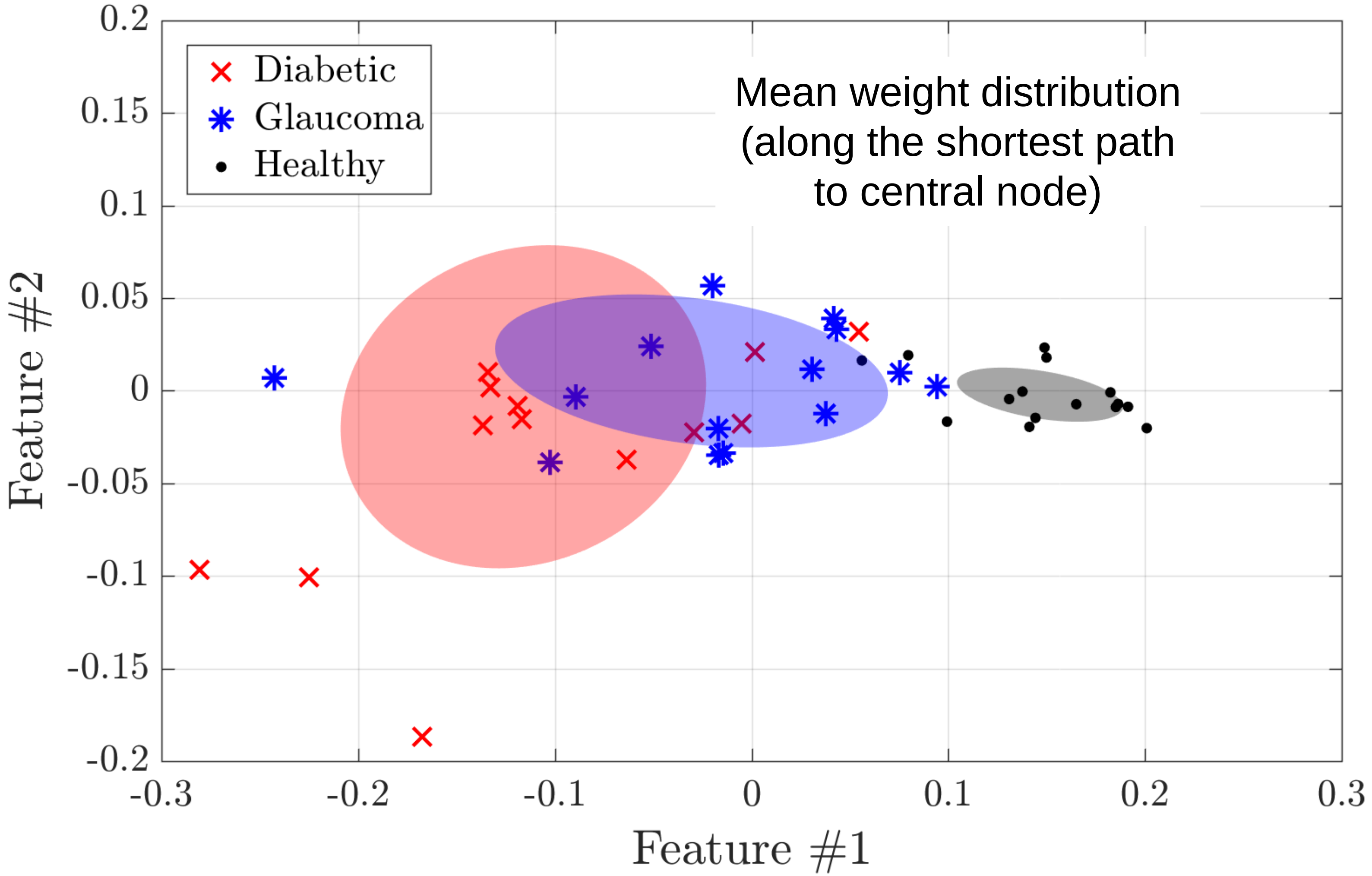
J. B. Tenenbaum et al, A global geometric framework for nonlinear dimensionality reduction. Science 290, 2319 (2000).

Performance of network features in manual segmentation:



P. Amil, F. Reyes-Manzano, L. Guzmán-Vargas, I. Sendiña-Nadal, C. Masoller, "*Network-based features for retinal fundus vessel structure analysis*", PLoS ONE 14, e0220132 (2019)

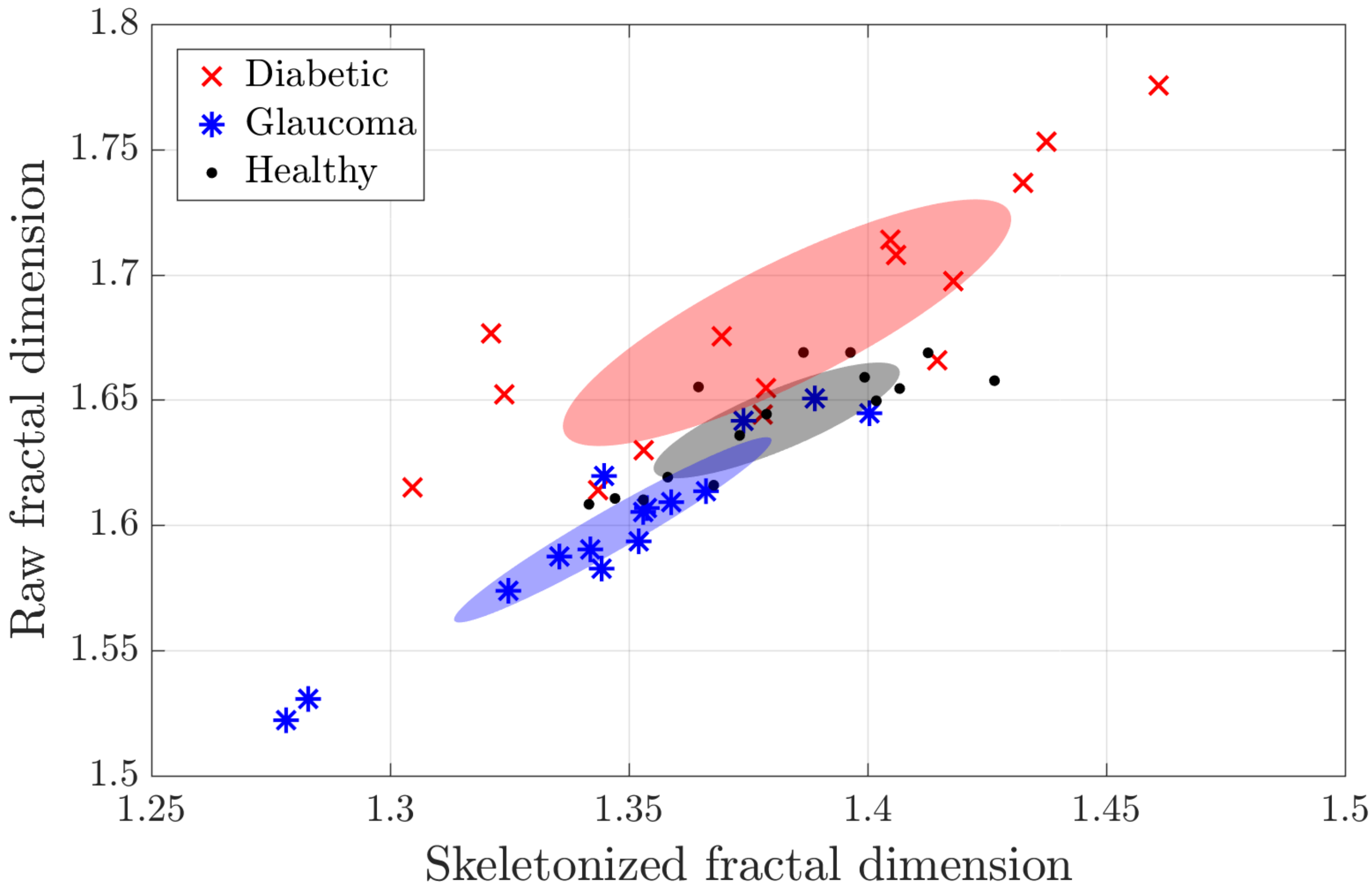
Performance of network features in manual segmentation:



In the automated segmentation:

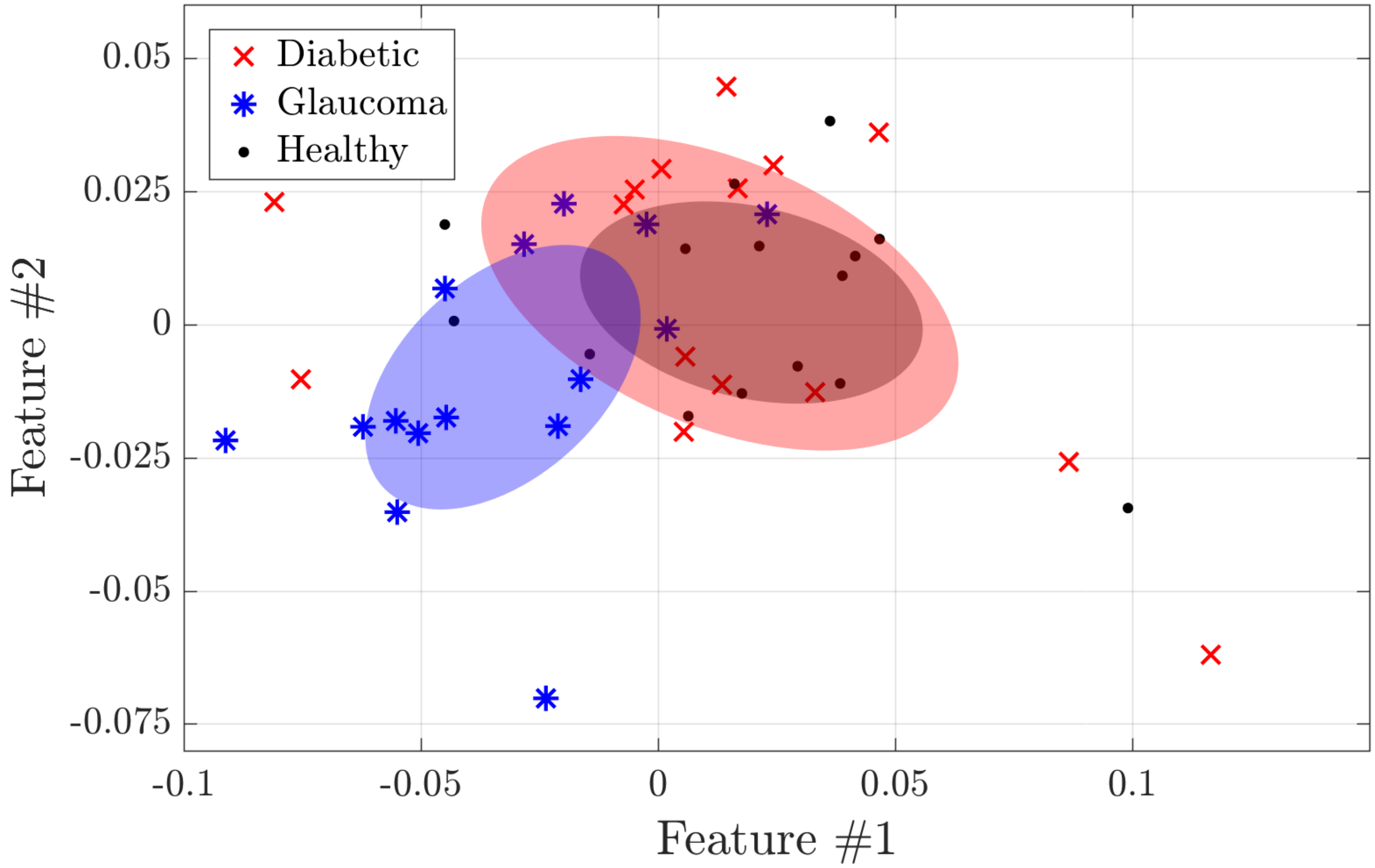
Fractal dimension analysis separates the three groups

$$D = \lim_{\varepsilon \to 0} \frac{\log(N(\varepsilon))}{\log(1/\varepsilon)}$$

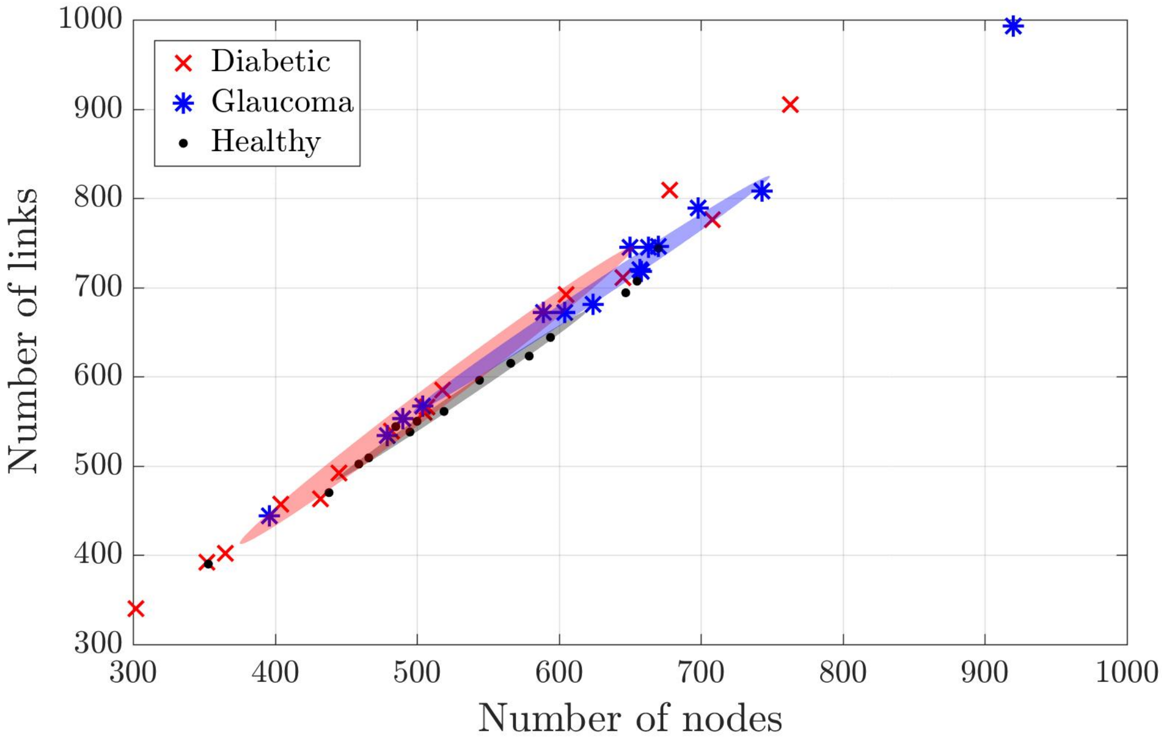


In the automated segmentation:

Mean weight distribution (along the shortest path to central node) identifies glaucoma



Simple network features do not differentiate



Summary I

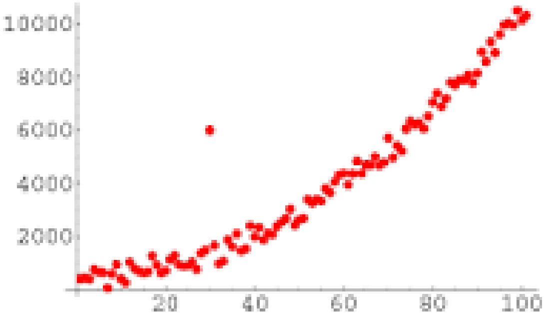
A measure for quantifying structural differences between graphs (different sizes, unlabeled nodes, undirected links) was used for

- EEG classification: two brain regions were identified that have different connection strength in control and in alcoholic subjects.

- Retina fundus image analysis: healthy and non-healthy patients were be distinguished by using high resolution images and manual segmentation.

What is an outlier?

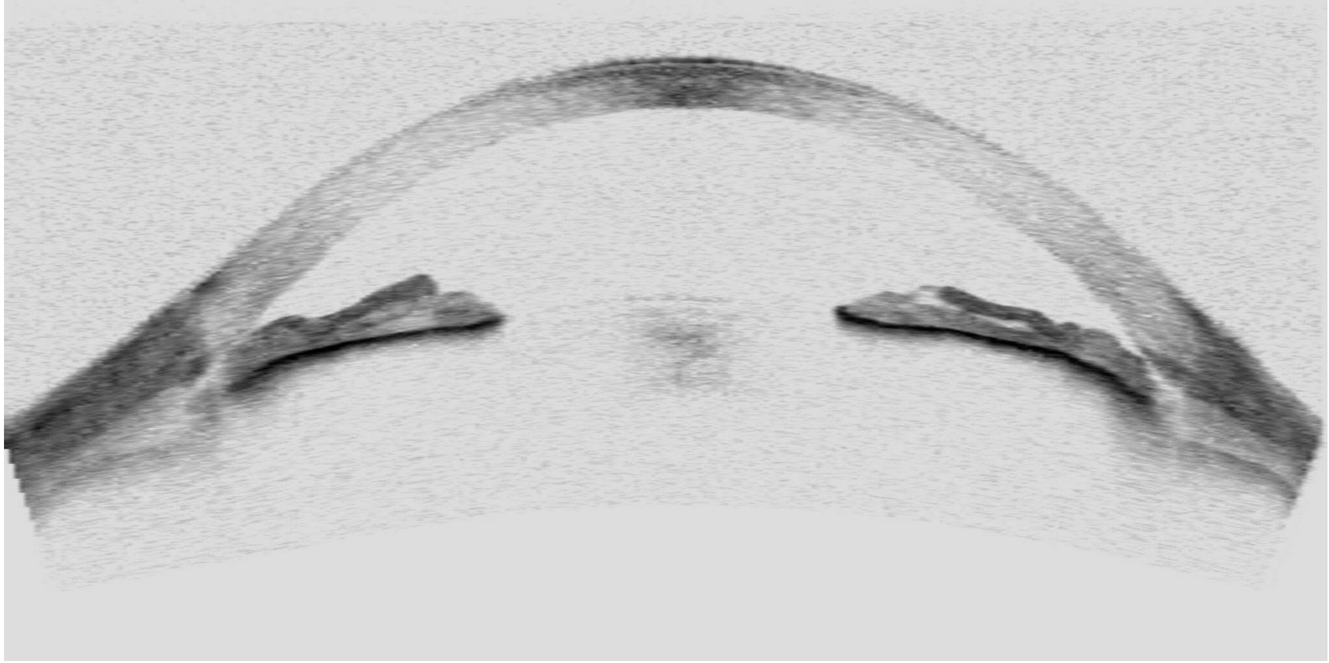


Practical definition: improved performance of machine learning algorithms when outliers are removed from the training set.

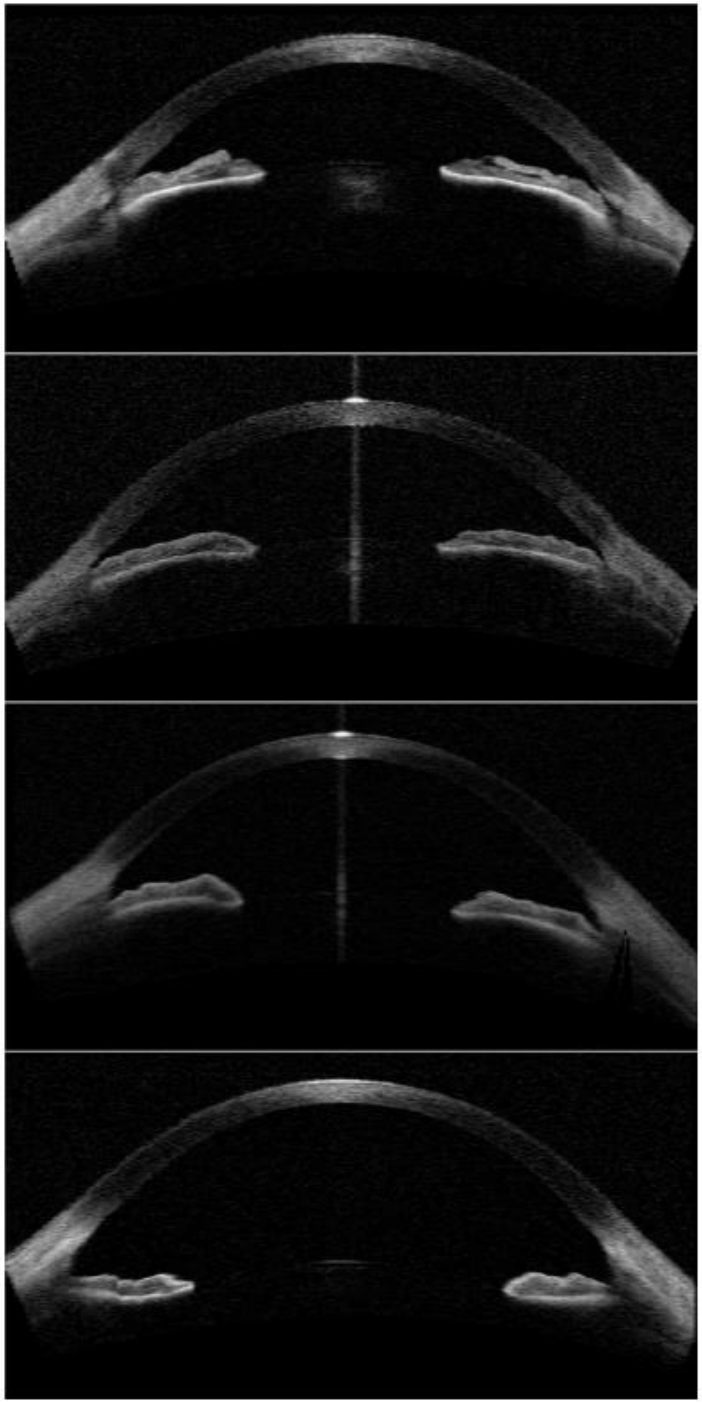
Motivation: unsupervised ordering of optical coherence tomography (OCT) images

- Used for diagnosis of glaucoma.
- More than 1000 images from patients of Instituto de Microcirugia Ocular (IMO, Barcelona).



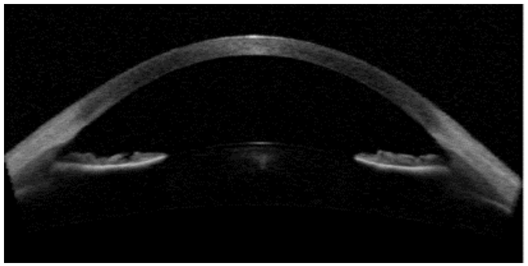
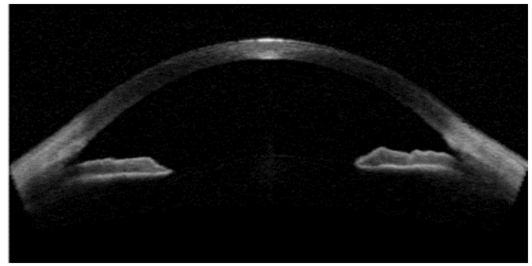
Classified into 4 categories by two ophthalmologists and a trained PhD student

- Closed
- Narrow
- Open
- Wide open

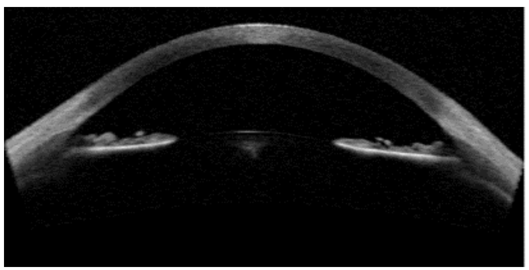
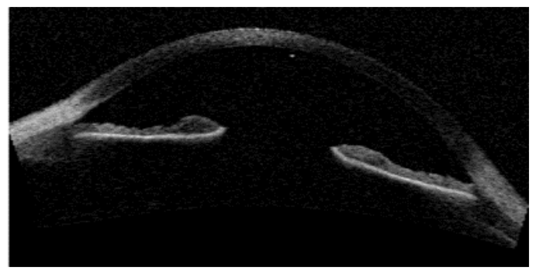


P. Amil et al., "*Unsupervised feature extraction of anterior chamber OCT images for ordering and classification*", Sci. Rep. 9, 1157 (2019).

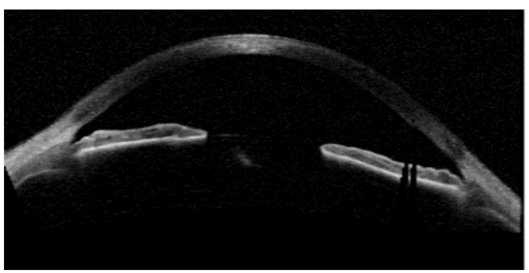
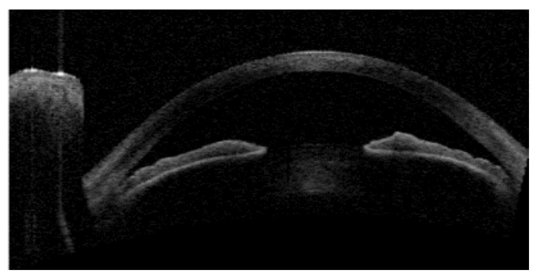
Using an appropriated distance (aligned Hellinger), by comparing pairs of images we extract features that can be used to order the images in a plane.

Small distance

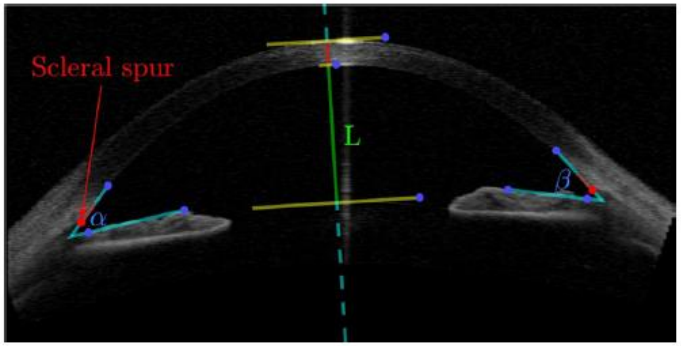
Medium distance

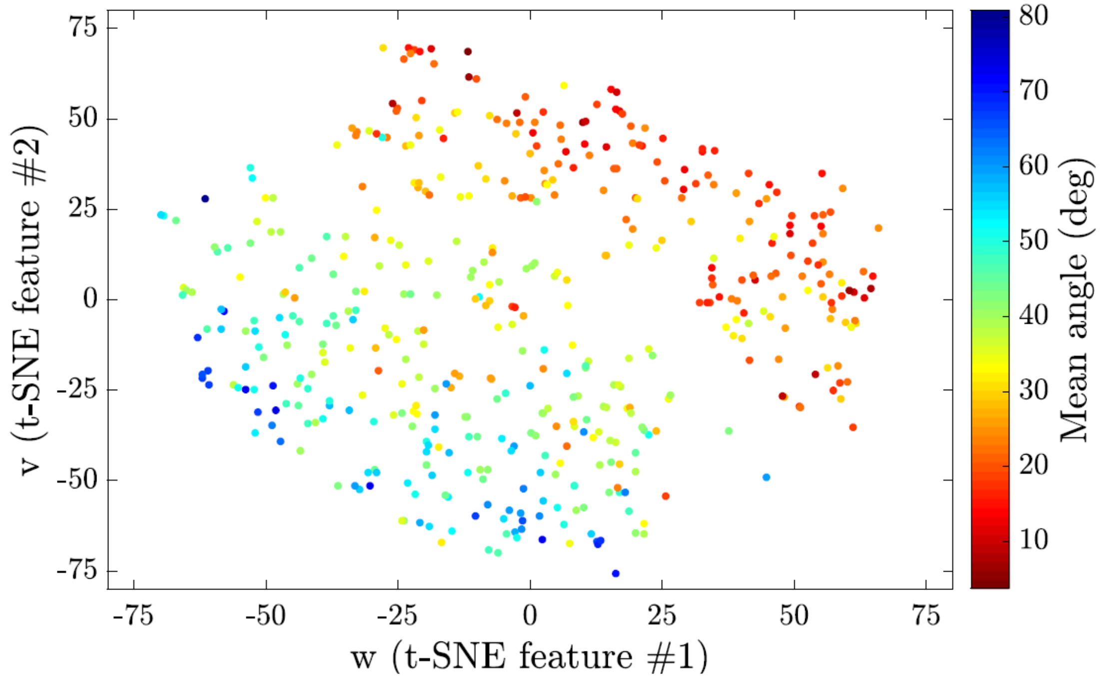



Large distance

P. Amil et al., “*Unsupervised feature extraction of anterior chamber OCT images for ordering and classification*”, Sci. Rep. 9, 1157 (2019).

Correlation between unsupervised features and the “manual” feature from expert annotation.





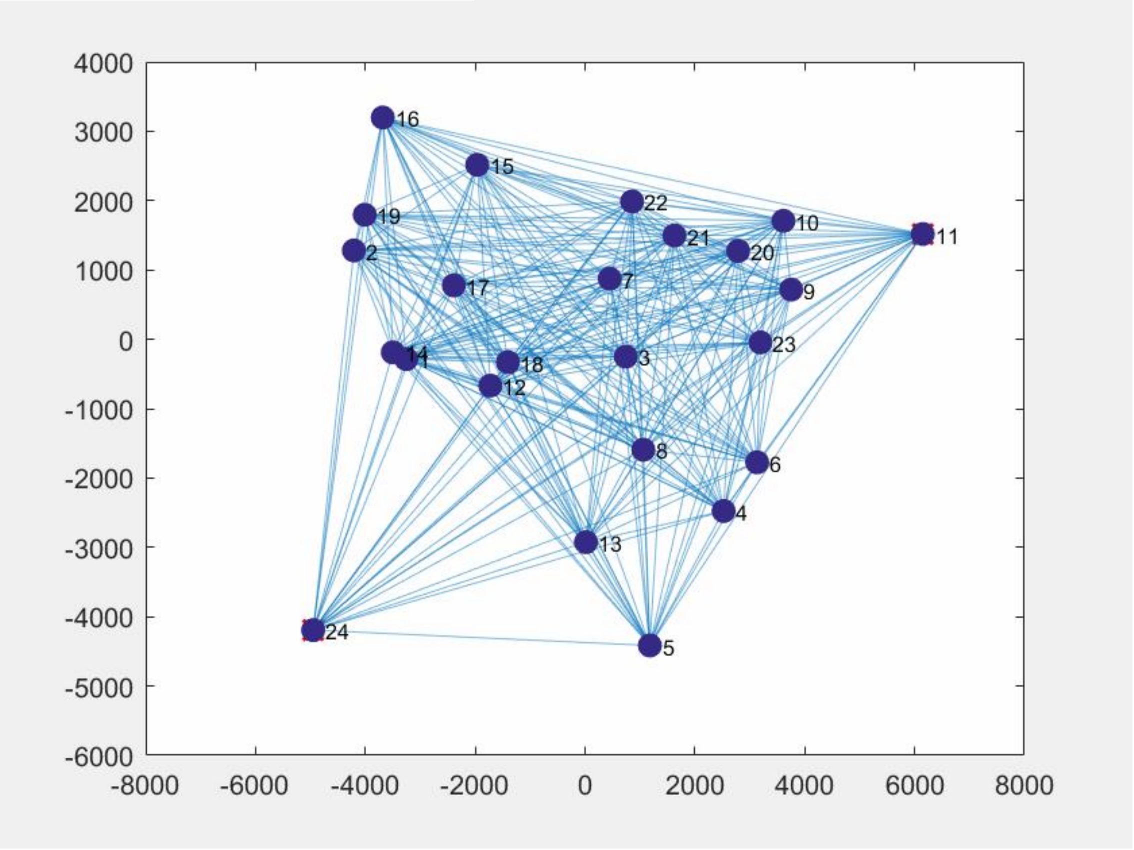
Can we improve this correlation if images with artifacts (outliers) are removed from the training set?

Two “network-based” & “distance based” outlier detection methods
- percolation
- manifold learning

First method: Outlier detection using percolation

$$V_i = \{v_1^i \ldots v_m^i\}$$ Feature vector describing each element of a dataset

$$D_{ij} = \left(\sum_k \left| v_k^i - v_k^j \right|^p \right)^{1/p}$$ Distance between any two elements of the dataset



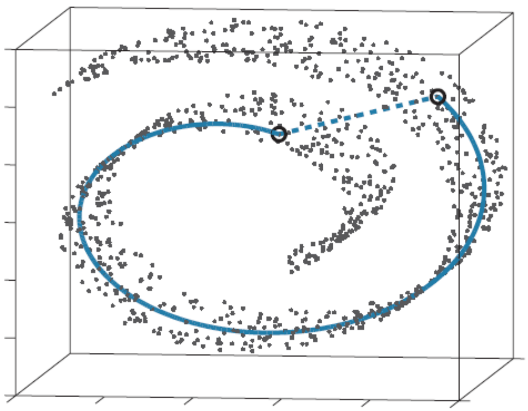
Outlier score = order in which elements disconnect from the giant component.

Parameter free.

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

Second method: Outlier detection using manifold learning

- *IsoMap* is a method for manifold learning. We define the outlier level with a measure of how well an element fits in the manifold.
- 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 $\mathbf{D'_{ij}}$
- $AL_i = 1 - {\rho_i}^2$
- Two parameters (integers):
  - Dimension of reduced space
  - # of geodesic neighbors



Distance in the high dimensional space (dash) and geodesic distance in the low-dimensional manifold.

Tenenbaum et al., A global geometric framework for nonlinear dimensionality reduction. Science 290, 2319-2323 (2000)

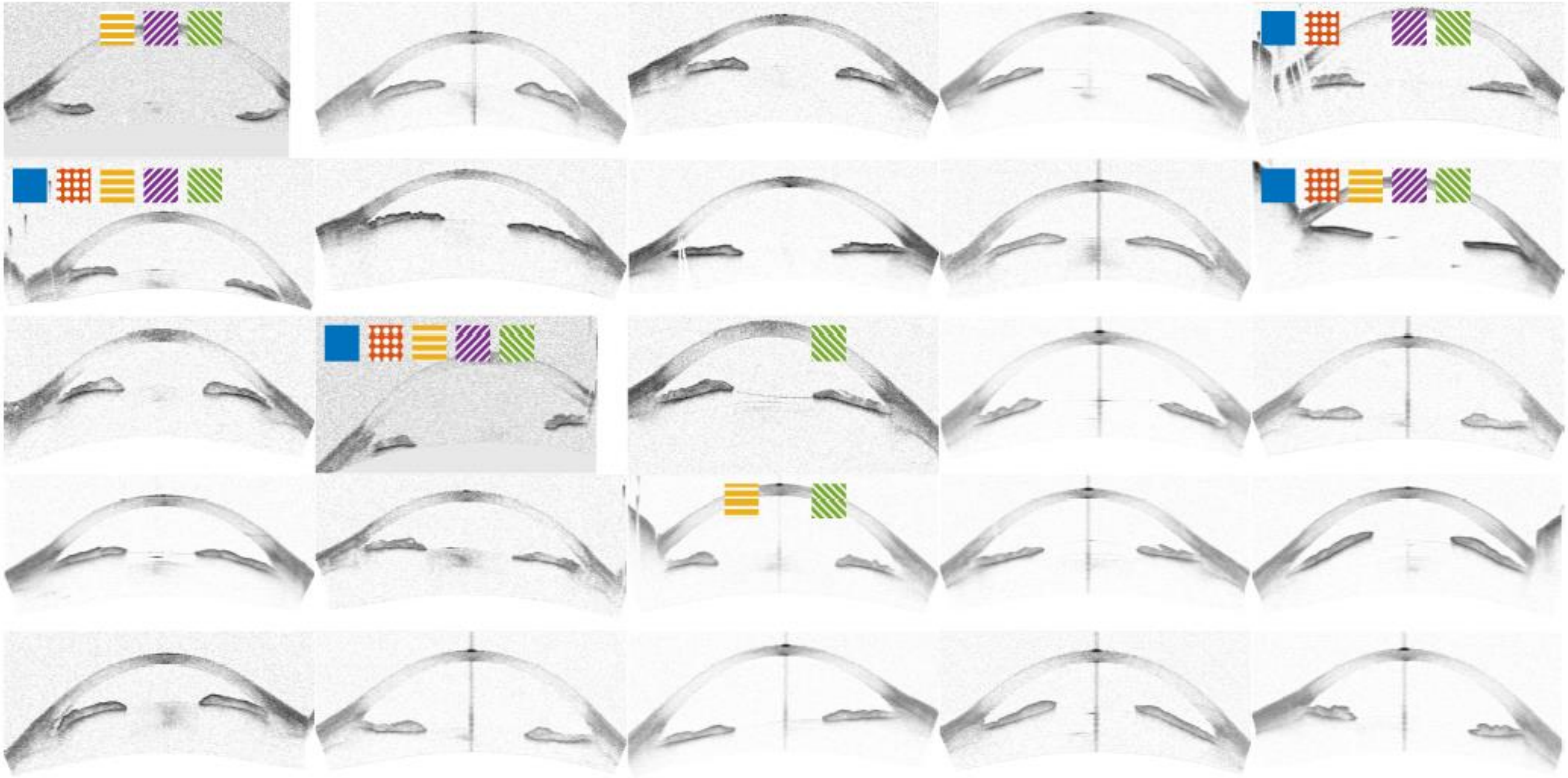
Comparison with other methods

- d2CM – Distance to center of mass. This method simple computes a “mean point” (center of mass) and computes the distance of each other point to this center of mass.
- Ramaswamy. A popular distance-based method. This method assigns an anomaly level to each point equal to its distance to its kth nearest neighbor.
- OCSVM - One Class Support Vector Machine. This method uses the inner product between the elements in the database to estimate a function that is positive in a subset of the input space where elements are likely to be found, and negative otherwise.

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.

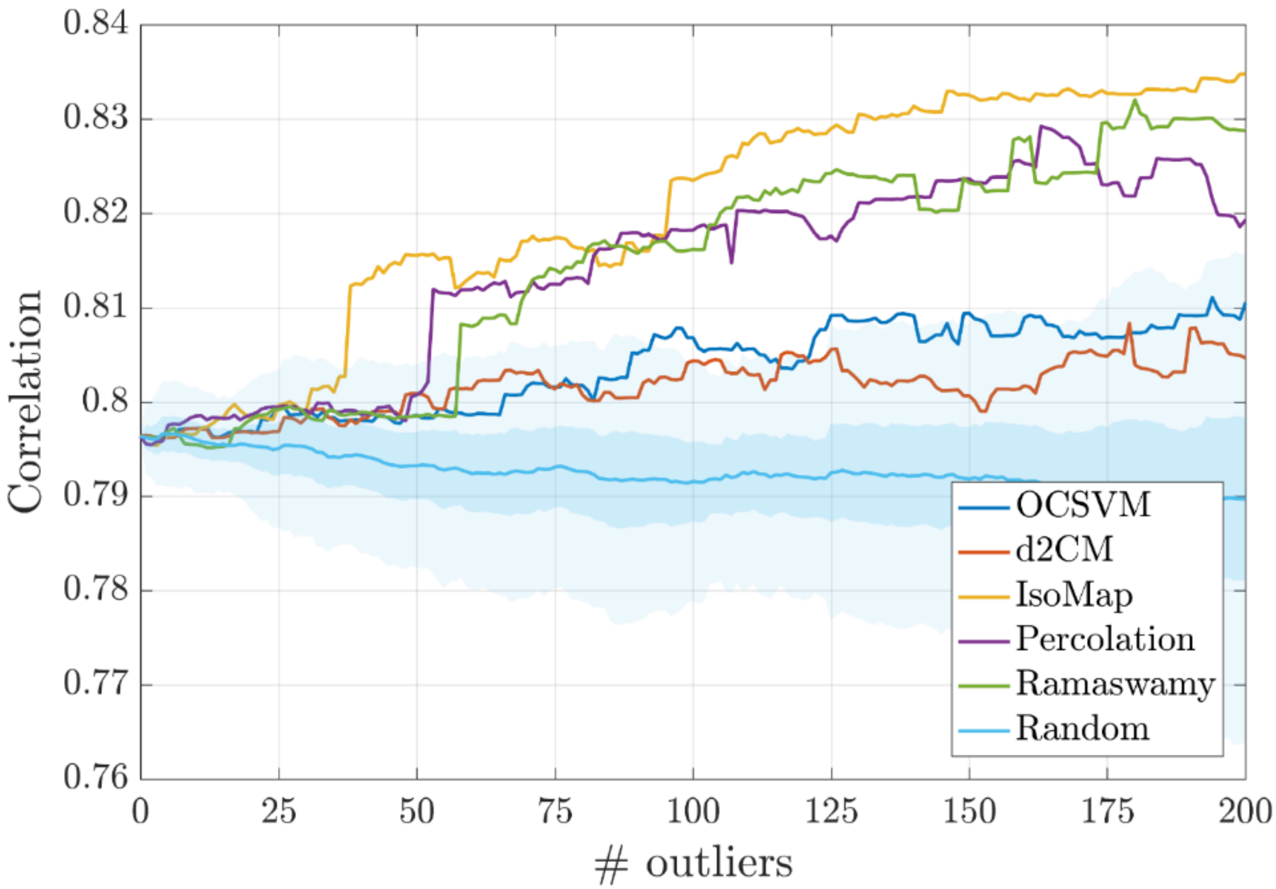
Results with OCT images (1/2)



All except the first one were randomly sampled. Marked images correspond to top 15% outlier score for OCSVM (Blue), distance to center of mass (Orange), IsoMap (Yellow), Percolation (Purple), and Ramaswamy (Green)

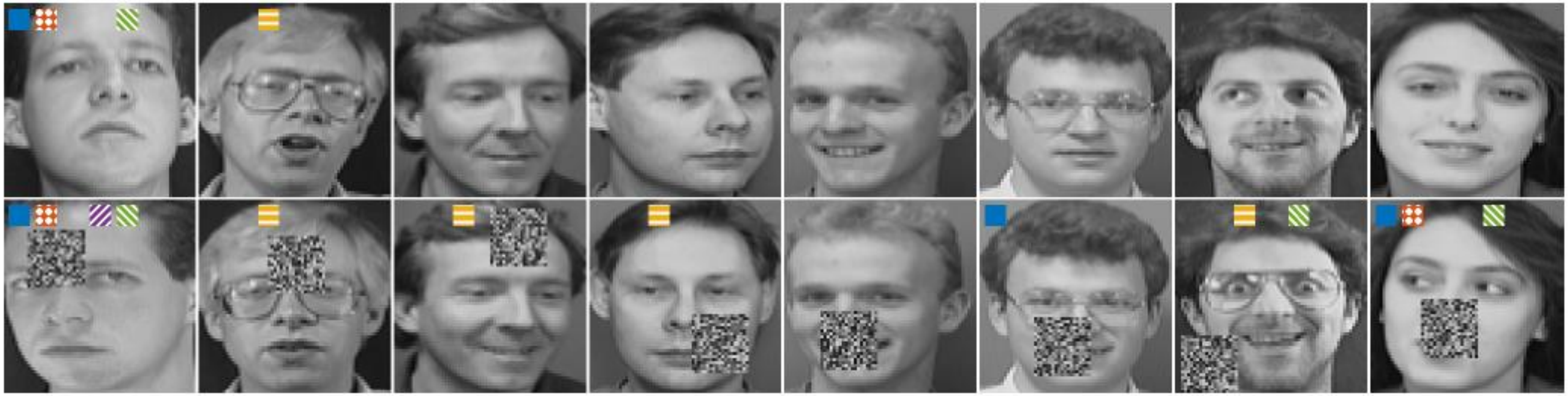
Results with OCT images (2/2)

Correlation between unsupervised features and the “manual” feature (mean angle) obtained from expert annotation.



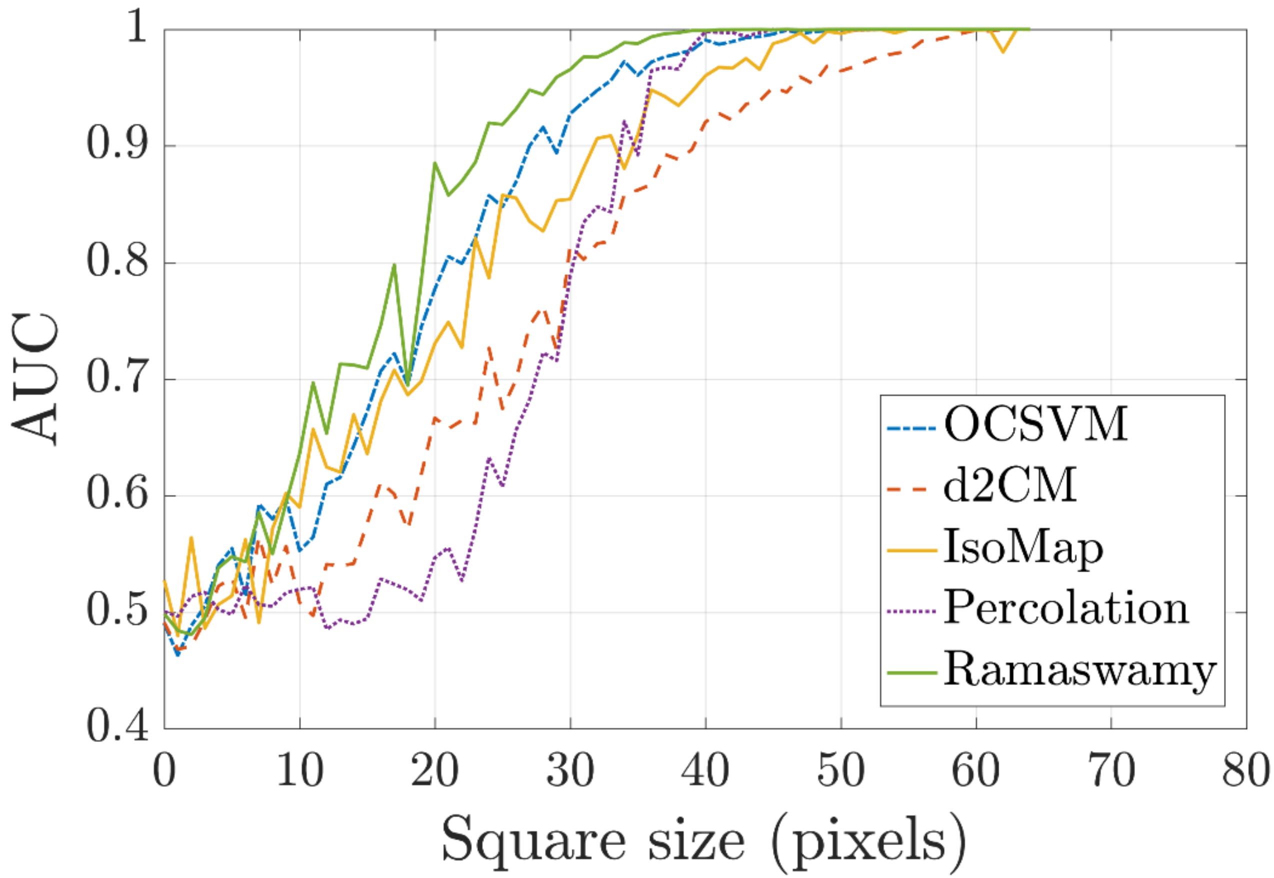
Can these methods work with other images?

- Freely available face database
- Added to some random images a square with gray-scale pixels whose color distribution is the same as that of the image.
- Measure success with the area under the ROC curve.



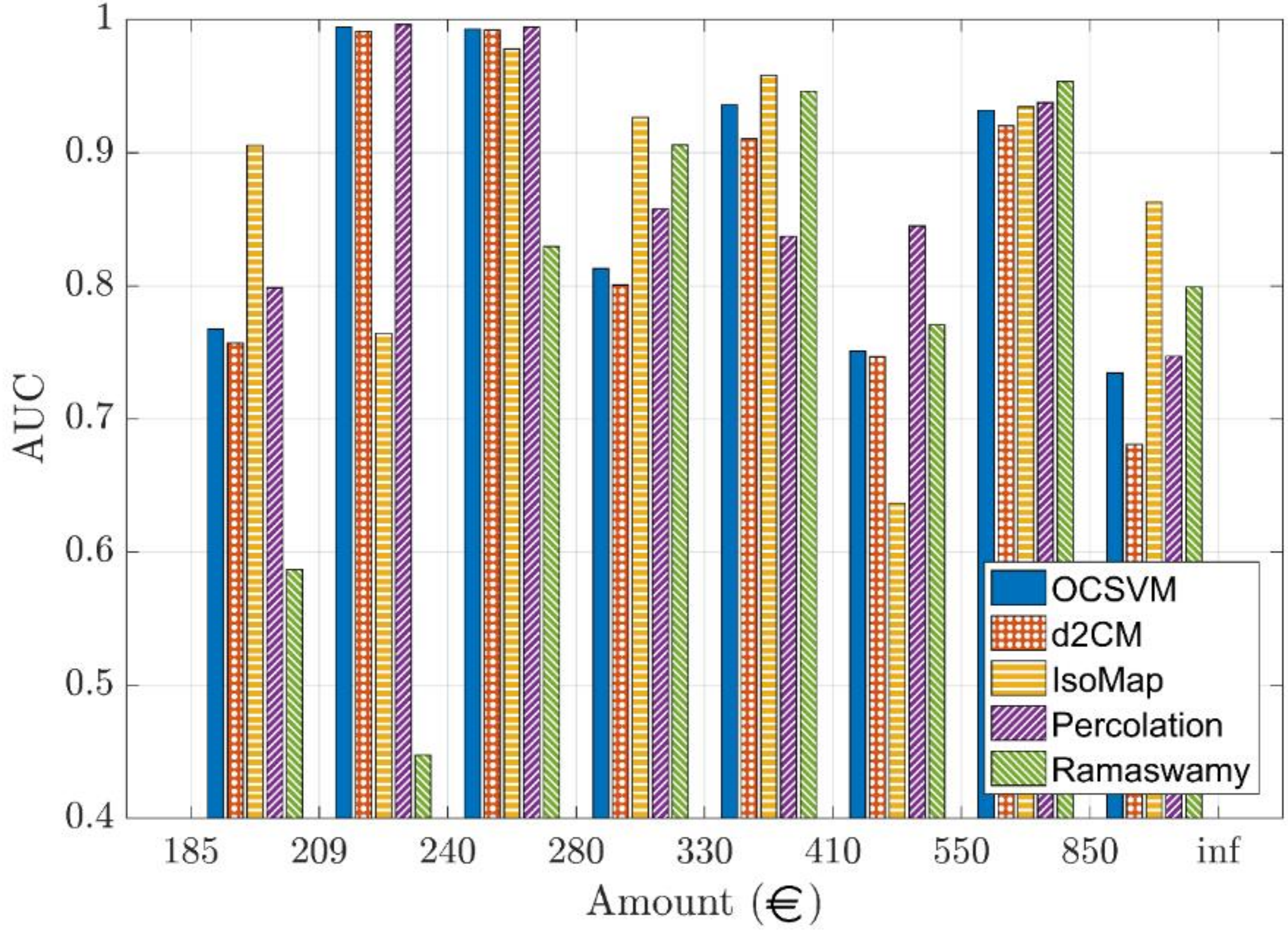
http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html

Results from the face database



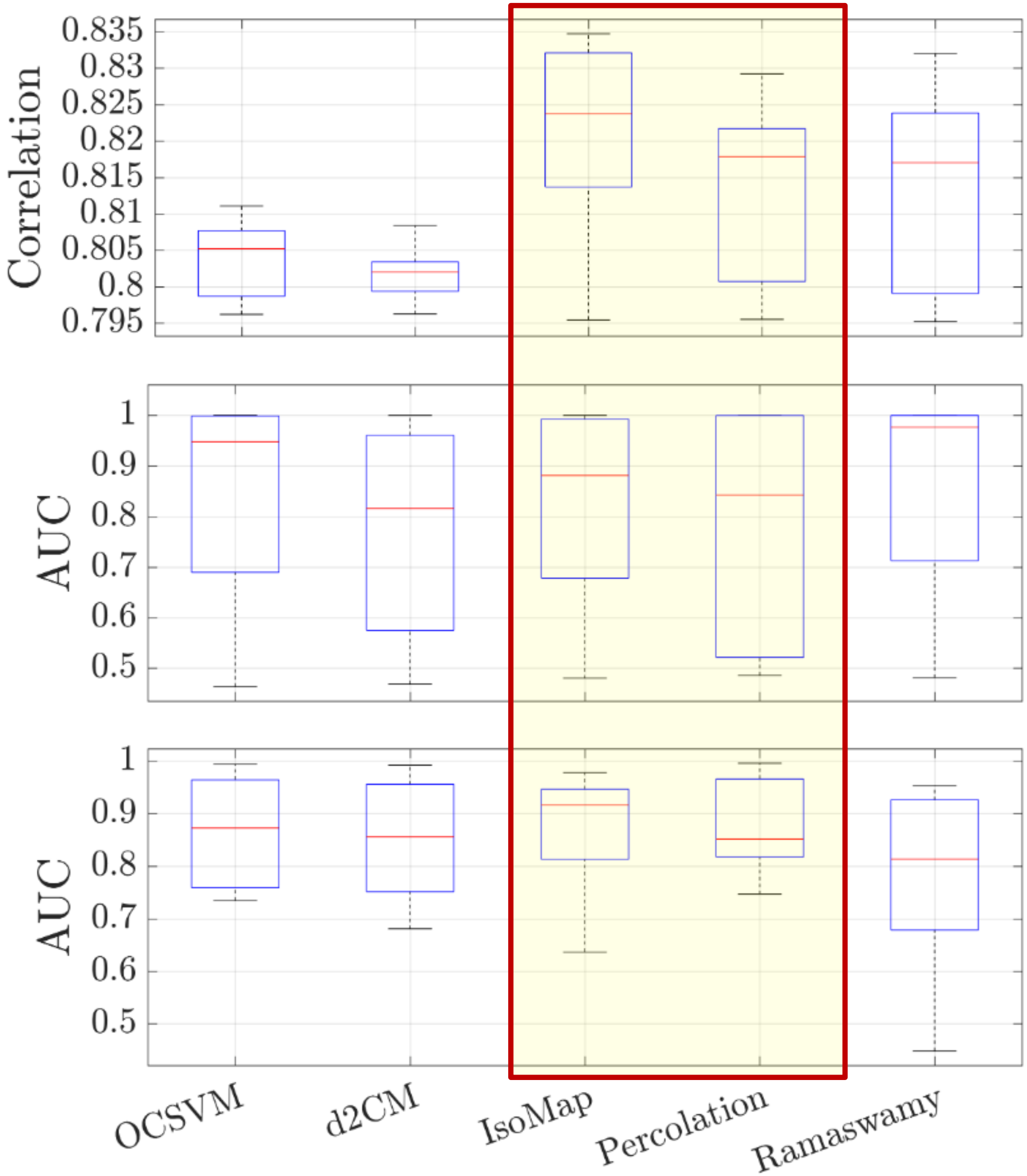
How about other types of elements?

Results from freely available credit card transactions (some identified as frauds)



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

Summary of results



OCT database

Face database

Credit card database

Summary II

- Two network-based methods to detect outliers perform well over different types of high-dimensional datasets, and their performance is competitive with other methods.
- The percolation method is parameter free, which makes it perfect for blind outlier finding.
- The *IsoMap* method has 2 parameters that when set properly can greatly outperform other methods, but is very sensitive to these parameters.

Work done in collaboration with

- T. A. Schieber, L. Carpi, A. Diaz-Guilera, P. M. Pardalos, M. G. Ravetti (Universidade Federal de Minas Gerais, University of Florida, University of Barcelona)
- Laura González, Elena Arrondo, Cecilia Salinas, Jose Luis Guell (Instituto de Microcirugia Ocular, Barcelona)
- Ulrich Parlitz (Max Plank Institute for Dynamics and Self-organization)
- Fabian Reyes-Manzano, Lev Guzman-Vargas (Instituto Politecnico Nacional, Mexico)
- Irene Sendiña-Nadal (Universidad Politecnica de Madrid)
- Nahuel Almeira (Universidad Nacional de Córdoba, Argentina)

Thank you for your attention !

- T. A. Schieber et al., "*Quantification of network structural dissimilarities*", Nat. Comm. 8, 13928 (2017).
- P. Amil et al., "*Unsupervised feature extraction of anterior chamber OCT images for ordering and classification*", Sci. Rep. 9, 1157 (2019).
- P. Amil et al., "*Network-based features for retinal fundus vessel structure analysis*", PLoS ONE 14, e0220132 (2019).
- P. Amil et al., "*Outlier mining methods based on graph structure analysis*", Front. Phys. 7, 194 (2019).