

# Combining complex networks and machine learning for image classification and outlier detection

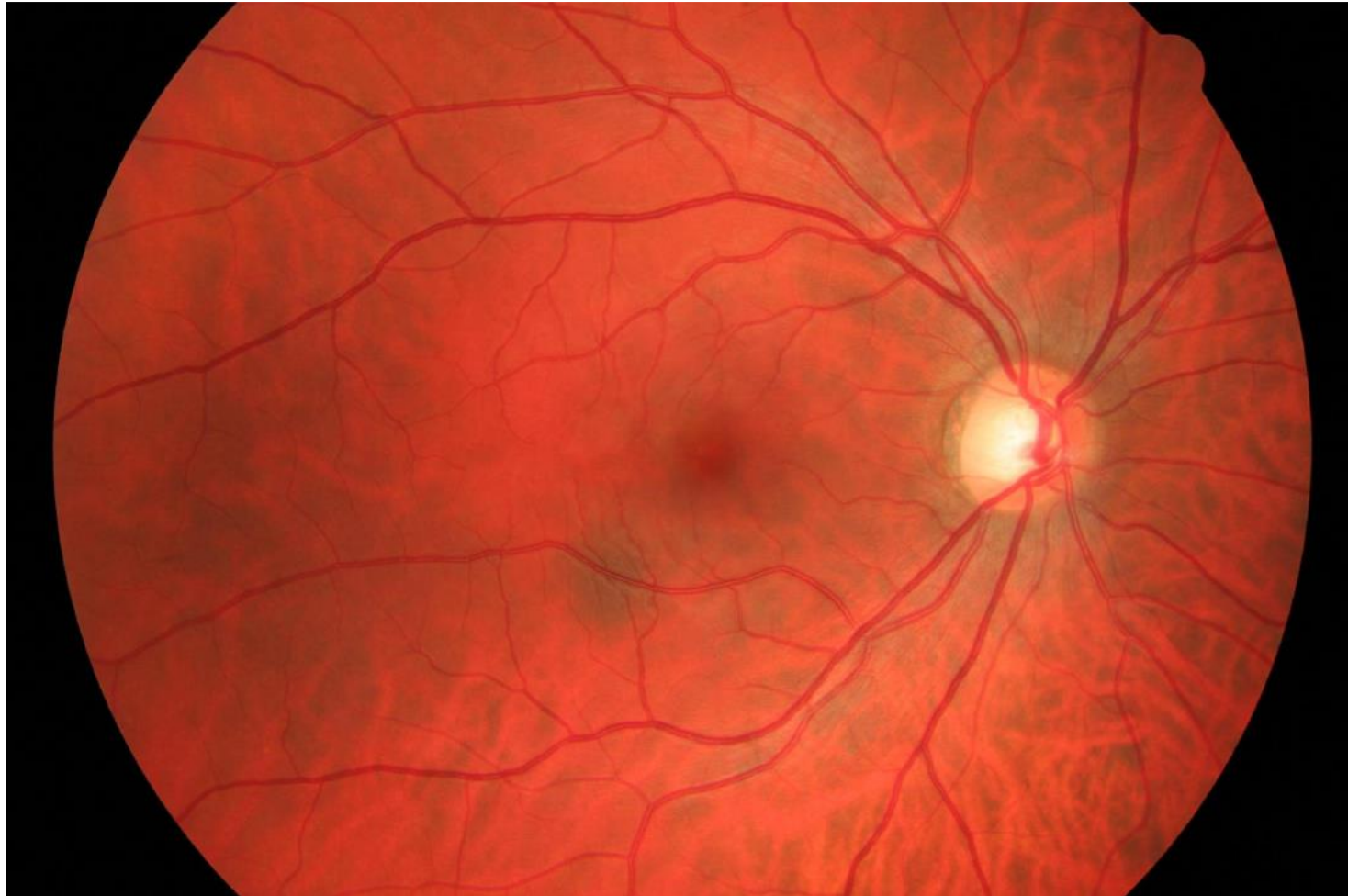
Pablo Amil and Cristina Masoller

Universitat Politècnica de Catalunya, Barcelona, Spain  
[www.fisica.edu.uy/~cris](http://www.fisica.edu.uy/~cris)



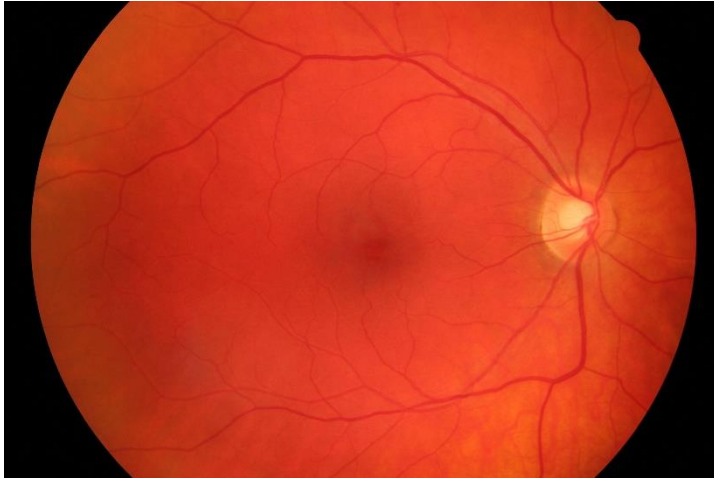
# Goal: use complex networks tools and machine learning algorithms to classify ophthalmic images

- Early diagnosis of eye deceases (glaucoma, diabetes) & follow up of treatments
- Person identification
- Signatures of other deceases (eg alterations in the brain arterial system)

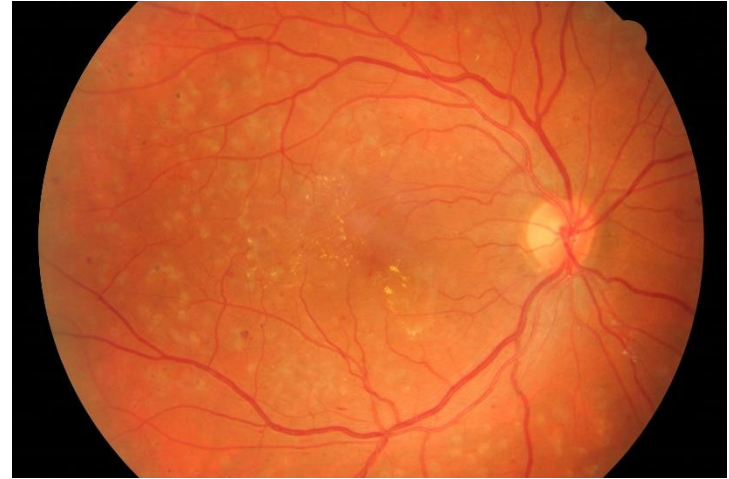


# Examples

Healthy



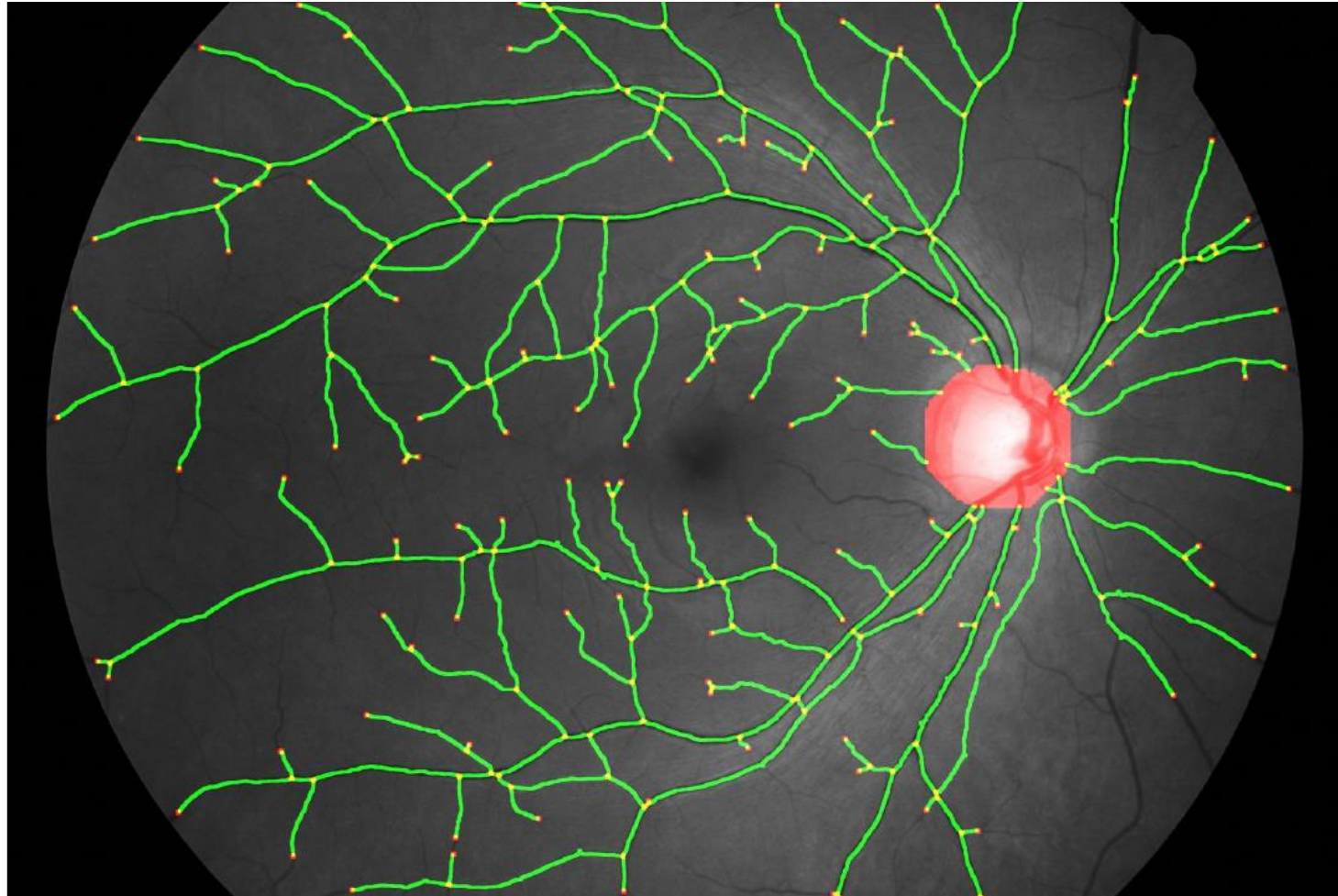
Diabetic



# Steps

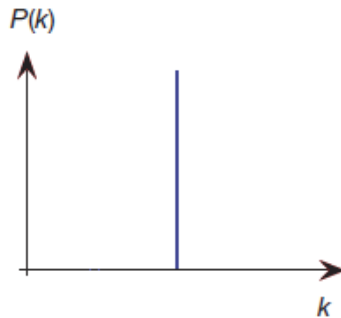
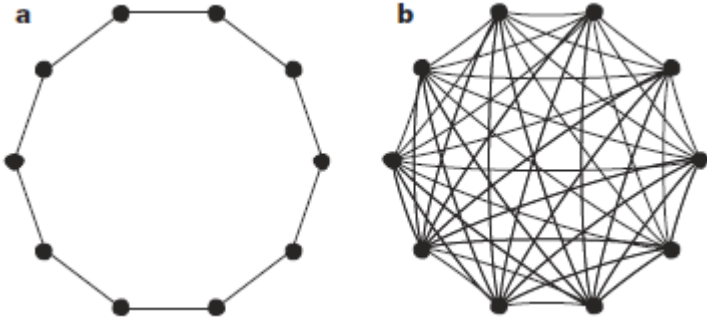
- Extract network
- Compare with other networks
- Use machine learning algorithms

⇒ Image classified

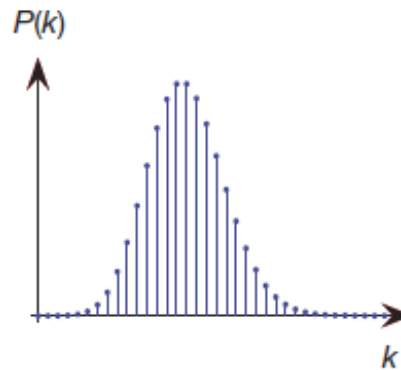
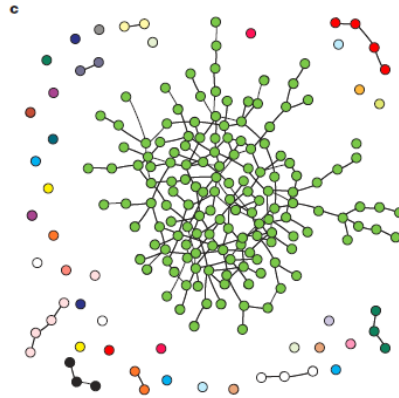


# How to compare *undirected* and *unweighted* networks with *different* number of nodes?

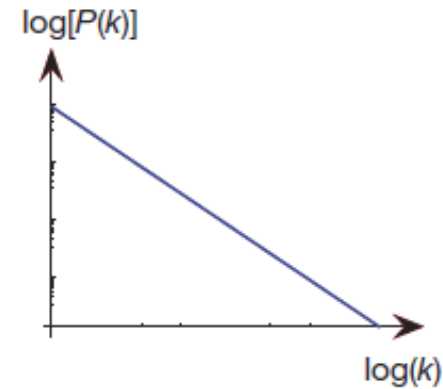
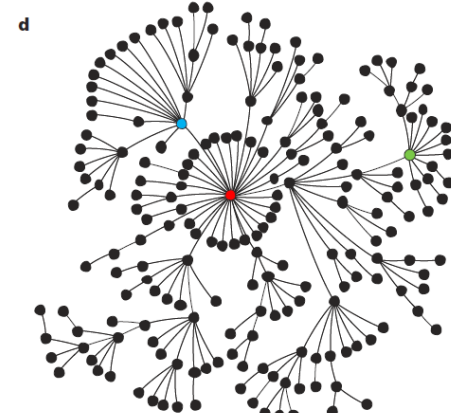
## Regular



## Random

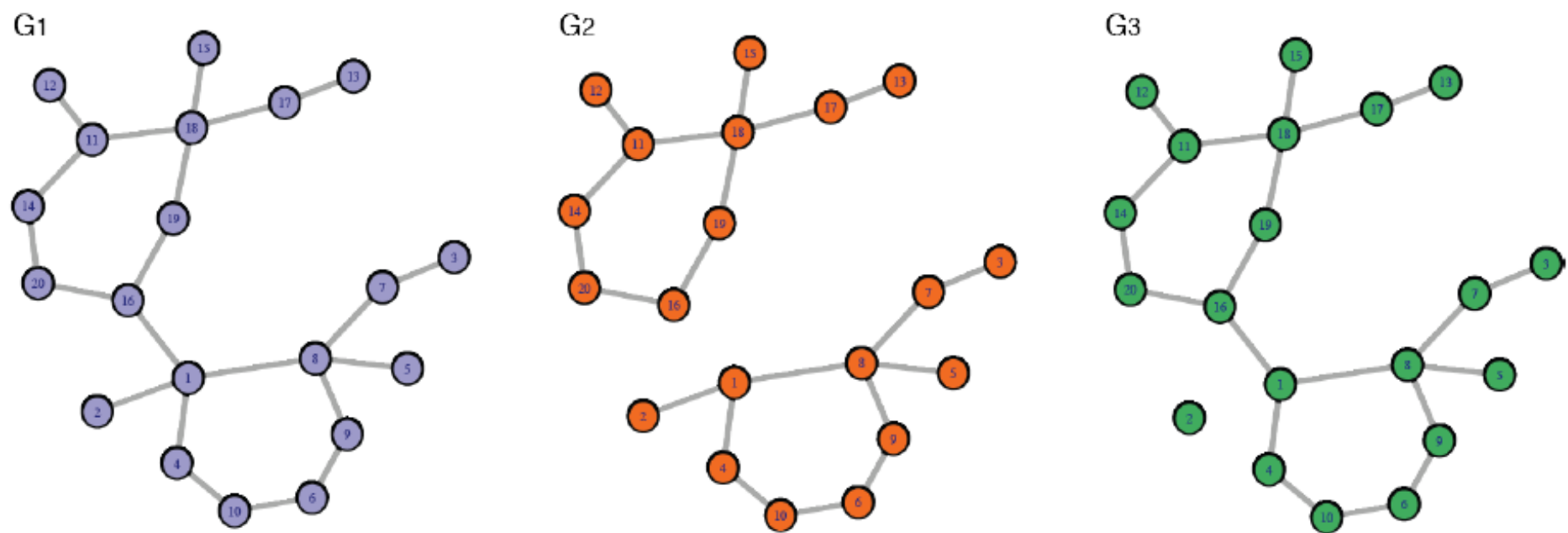


## Scale-free

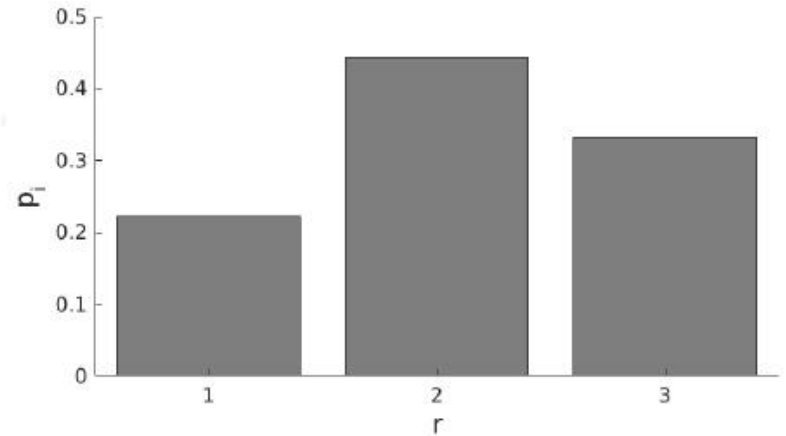
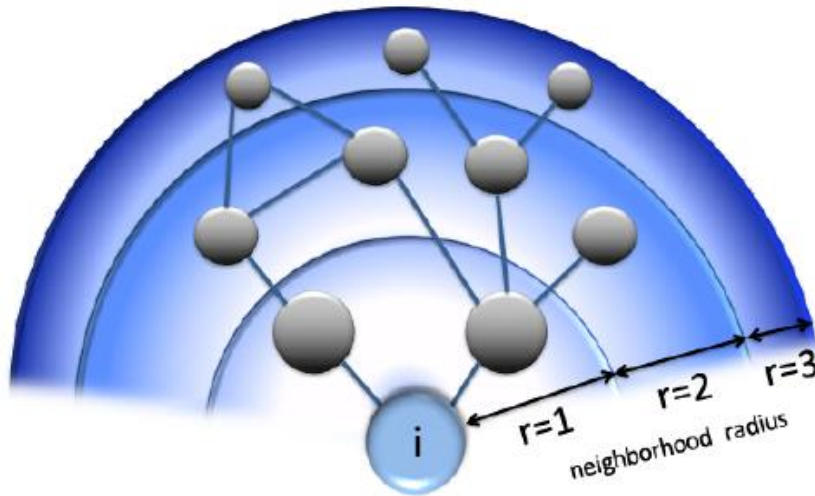


# To detect structural differences between networks we need a precise measure to compare them.

- Degree, centrality, assortativity, betweenness, etc. distributions provide *partial* information.
- Main problem: not all the links have the same importance.



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



A network with  $N$  nodes is characterized by  $N$  “node distance distributions”

$$\{p_1, p_2, \dots, p_N\}$$

If two networks have the same set of pdfs  $\Rightarrow$  they have the same diameter, average path length, etc.

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

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

average node distance distribution:  $\mu = \langle p_i \rangle_i$

Kullback distance between

$$p_i \text{ and } \mu: J(p_i, \mu) = \sum_r p_i(r) \log \left( \frac{p_i(r)}{\mu(r)} \right)$$

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



# Dissimilarity measure

$$D(G, G') = w_1 \sqrt{\frac{\mathcal{J}(\mu_G, \mu_{G'})}{\log 2}} + w_2 \left| \sqrt{\text{NND}(G)} - \sqrt{\text{NND}(G')} \right| \quad w_1=w_2=0.5$$

compares the  
averaged  
connectivity

compares the  
heterogeneity of the  
connectivity distances

- Extensive numerical experiments demonstrate that isomorphic networks return  **$D=0$** .
- Computationally efficient.
- Allows to compare networks with different number of nodes.

*T. A. Schieber et al, Nat. Comm. 8, 13928 (2017)*

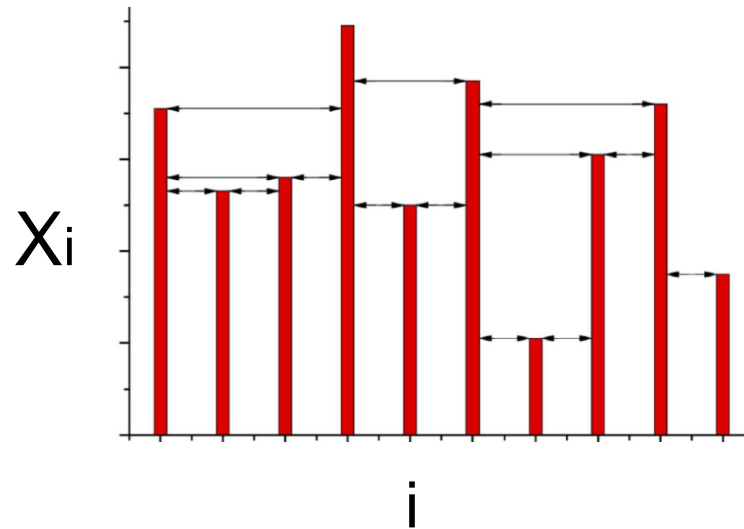
**First example of application:  
classification of EEG signals**

# 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 time series is transformed into a network using the horizontal visibility rule.

\* <https://archive.ics.uci.edu/ml/datasets/eeg+database>

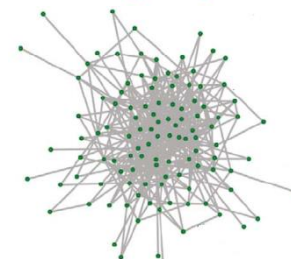
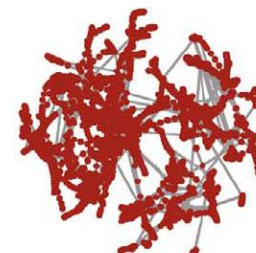
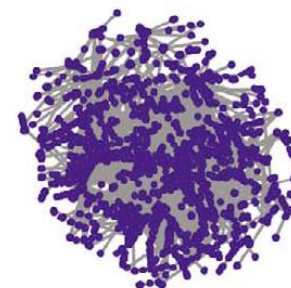
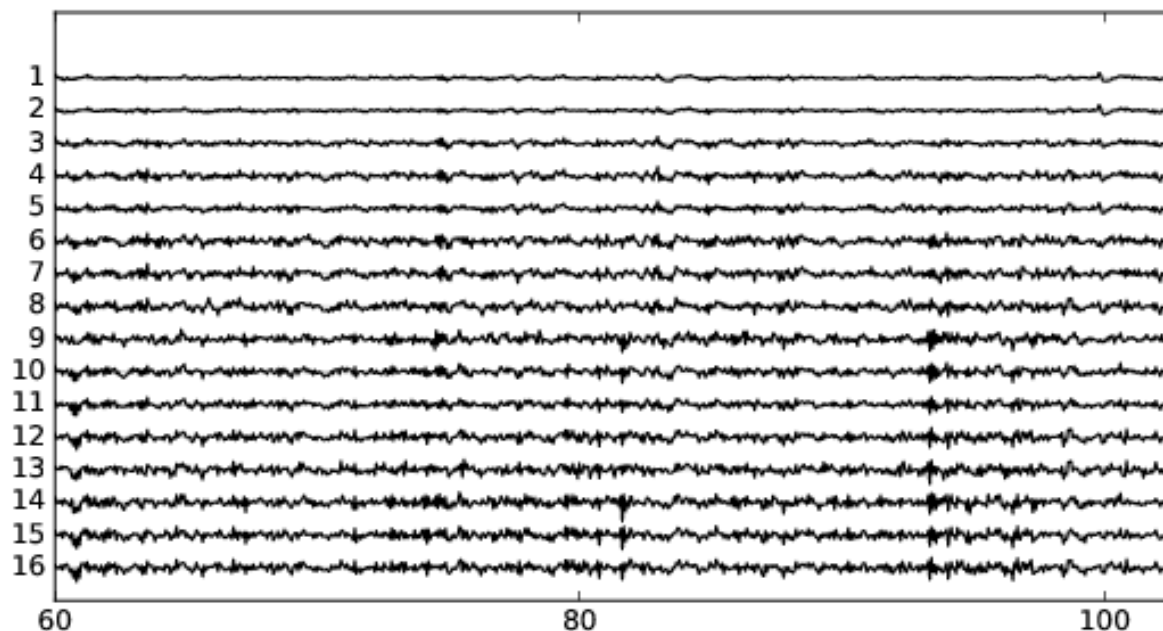
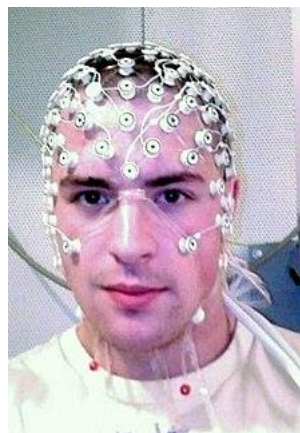
# Horizontal visibility graph



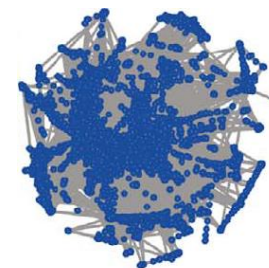
- Data points  $i$  and  $j$  are connected if there is “visibility” between them.
- We obtain an unweighted and undirected graph.
- Parameter free!

*Luque et al PRE (2009); Gomez Ravetti et al, PLoS ONE (2014)*

# For each subject dataset has 64 channels $\Rightarrow$ 64 networks

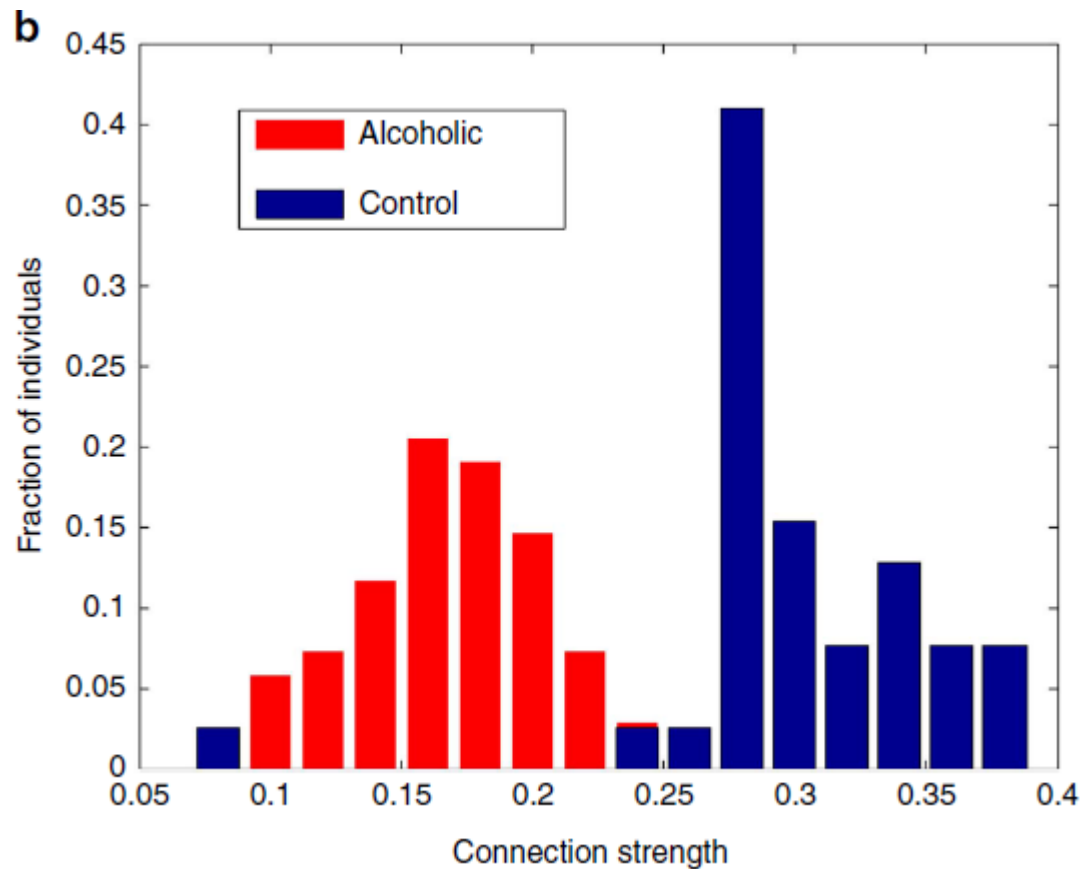


...



- The connection strength between two brain regions is:  $1-D(G,G')$

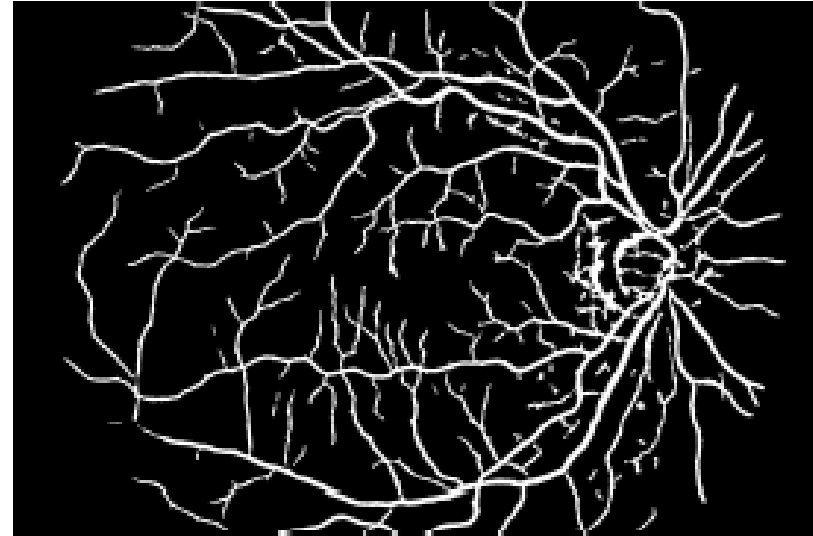
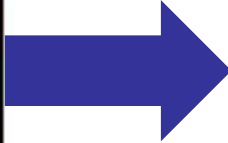
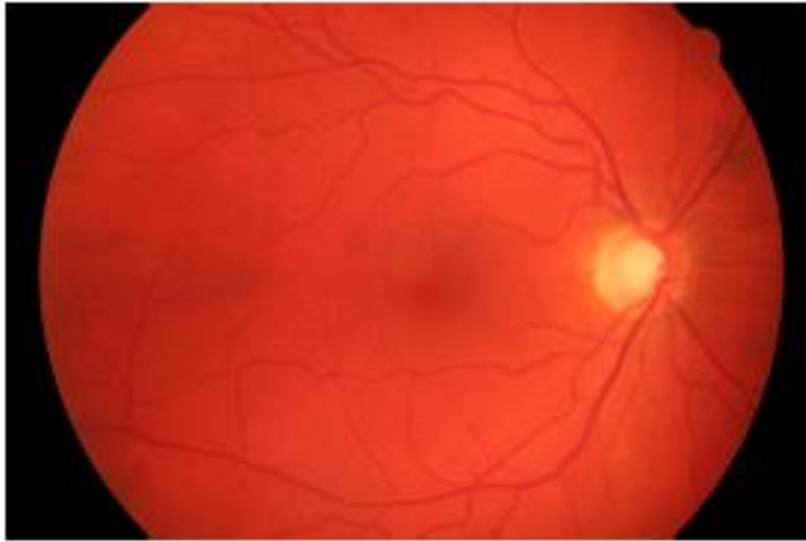
**We identified two brain regions (called ‘nd’ and ‘y’), where the connection strength between these regions is higher in control than in alcoholic subjects.**



*T. A. Schieber et al, Nat. Comm. 8, 13928 (2017)*

**Second application:  
classification of retina images**

# Segmentation

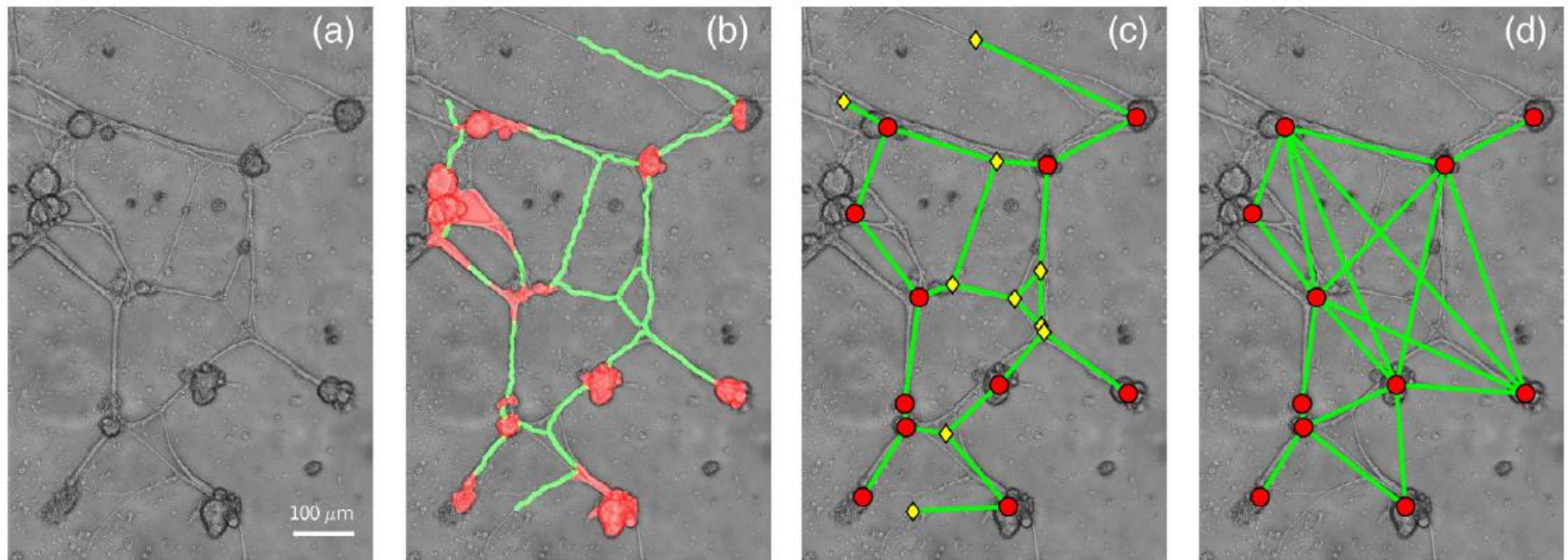


Problem: arteries and veins networks collapse.



# Unsupervised segmentation algorithm adapted from an algorithm developed for segmenting images of cultured neuronal networks.

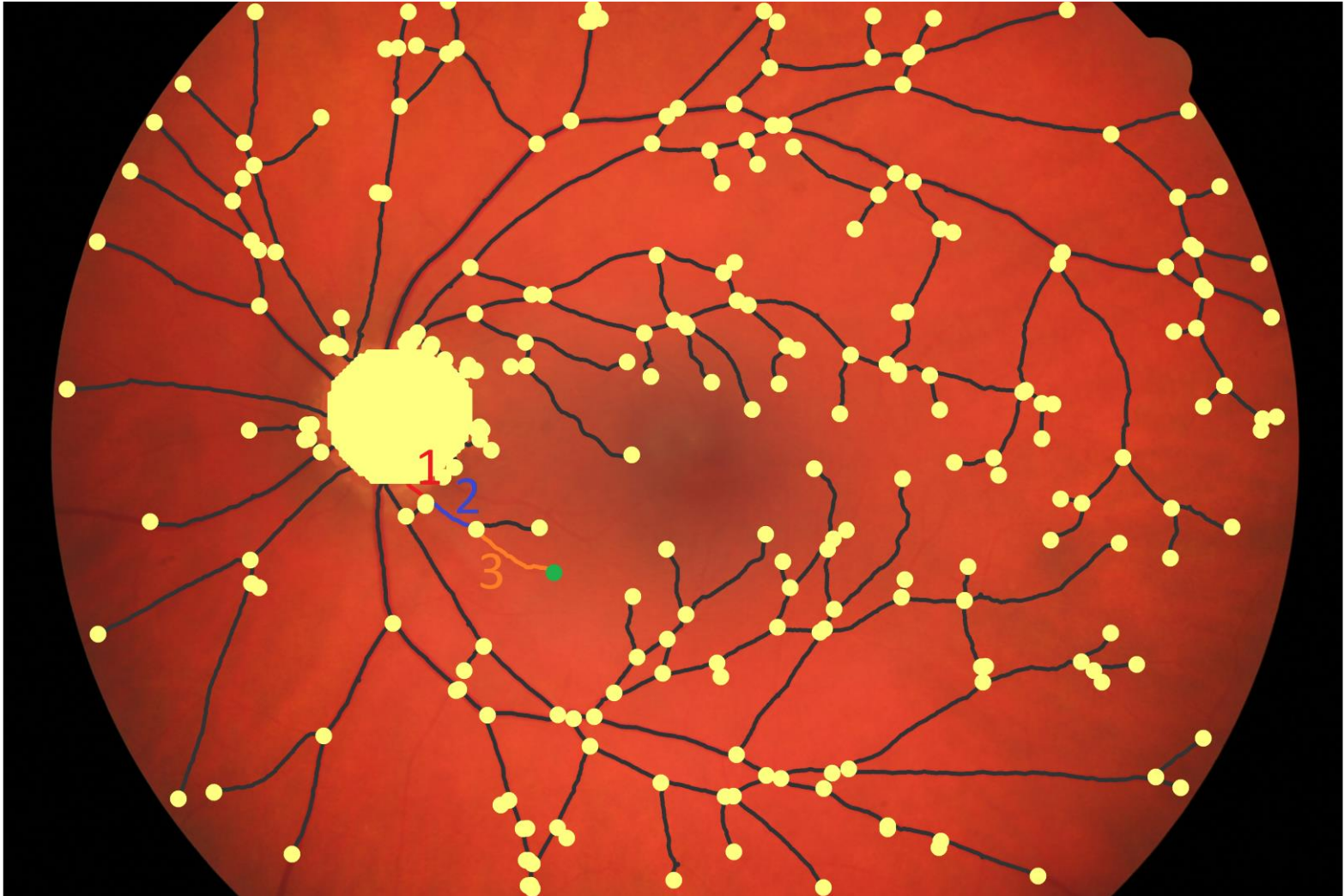
A. Tlaie, L.M. Ballesteros-Esteban and I. Leyva et al. / *Chaos, Solitons and Fractals* 119 (2019) 284–290



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).

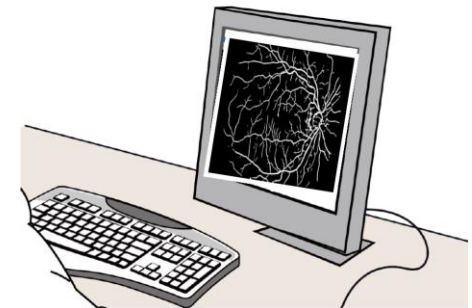
**Identification of nodes and links.**

**Then, analysis of the connectivity paths to the central node (optical nerve)**



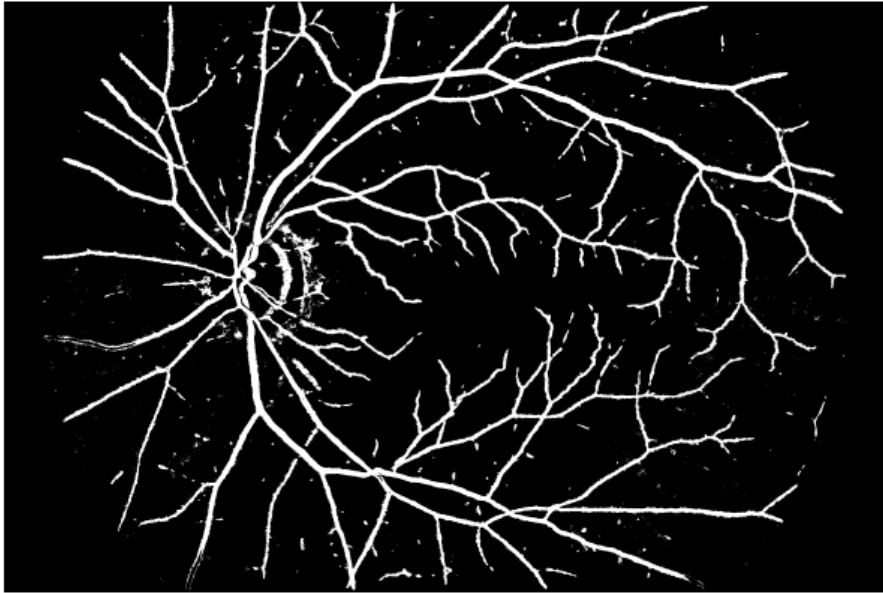
# Data (1/2)

- High-resolution public database with
  - 15 healthy subjects
  - 15 glaucoma
  - 15 diabetic retinopathy
- For every subject there is
  - fundus photography
  - manual segmentation of the vessels done by an expert.
- From each fundus photography  $\Rightarrow$  automated segmentation.

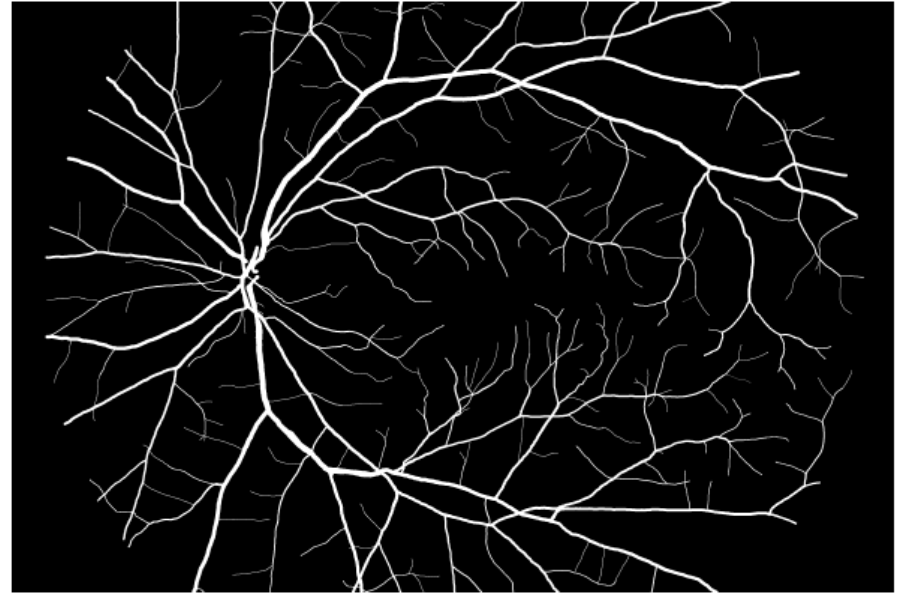


# Comparison

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

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

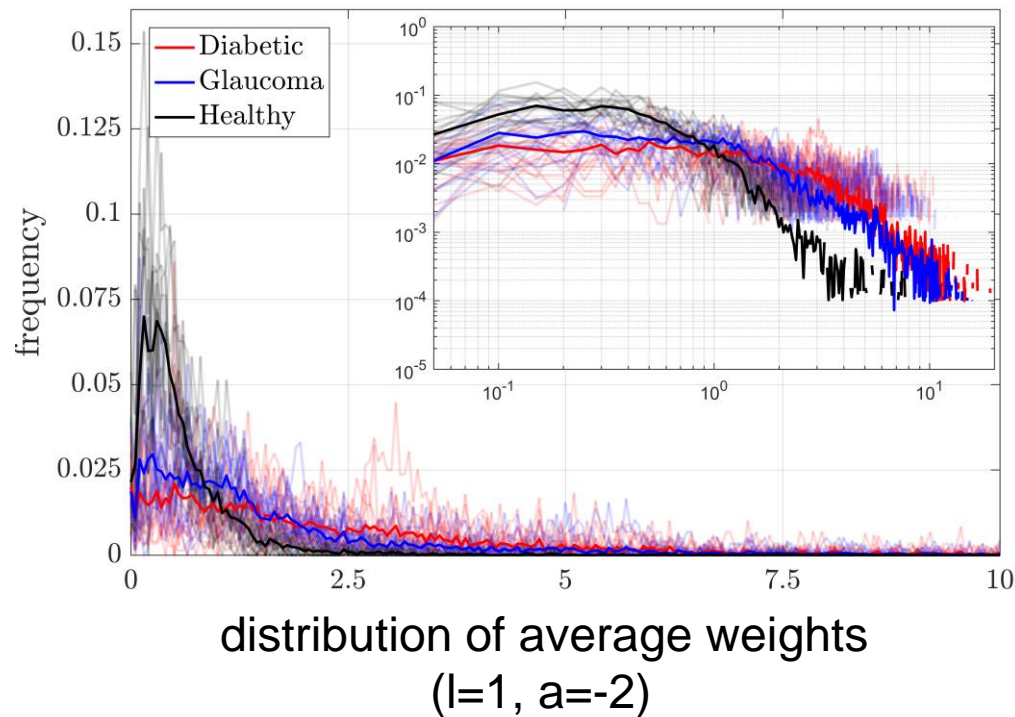
length and width (in # of pixels) of the segment that connects nodes i and j

- 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 central node
- Distribution of weighted degrees



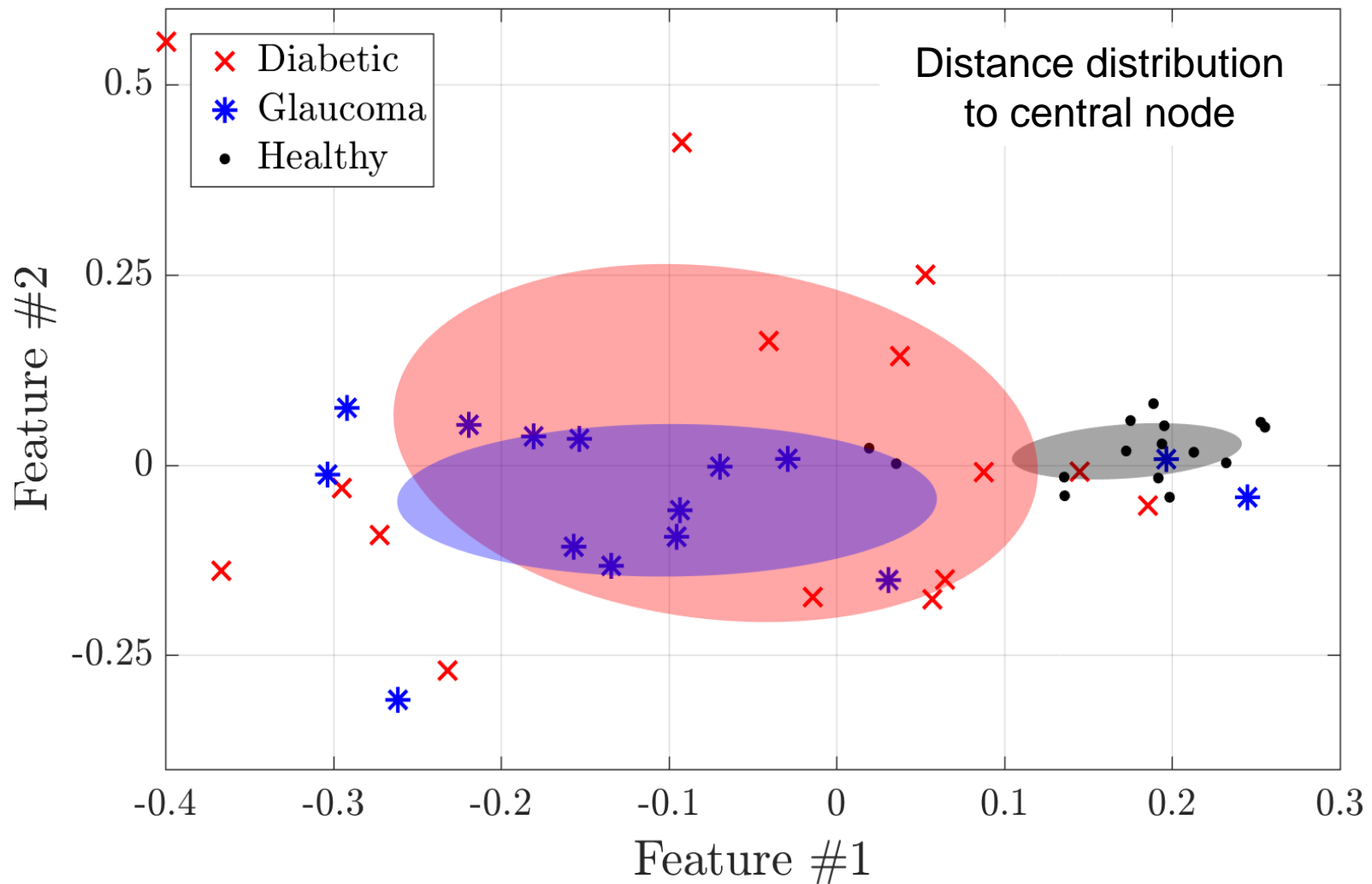
## Methodology (2/2)

- We use the Jensen-Shannon (JS) divergence to compare distributions (image  $i$  with all other images)
- For each image  $i$  we obtain a vector
$$\{d_{i1}, d_{i2}, \dots, d_{iN}\} \text{ (} N = \text{number of images)}$$
whose elements are the distances between the distributions extracted from image  $i$  and image  $j$  ( $j$  in  $1 \dots N$ ).
- This vector of  $N$  “features” characterizes image  $i$ .
- 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).*

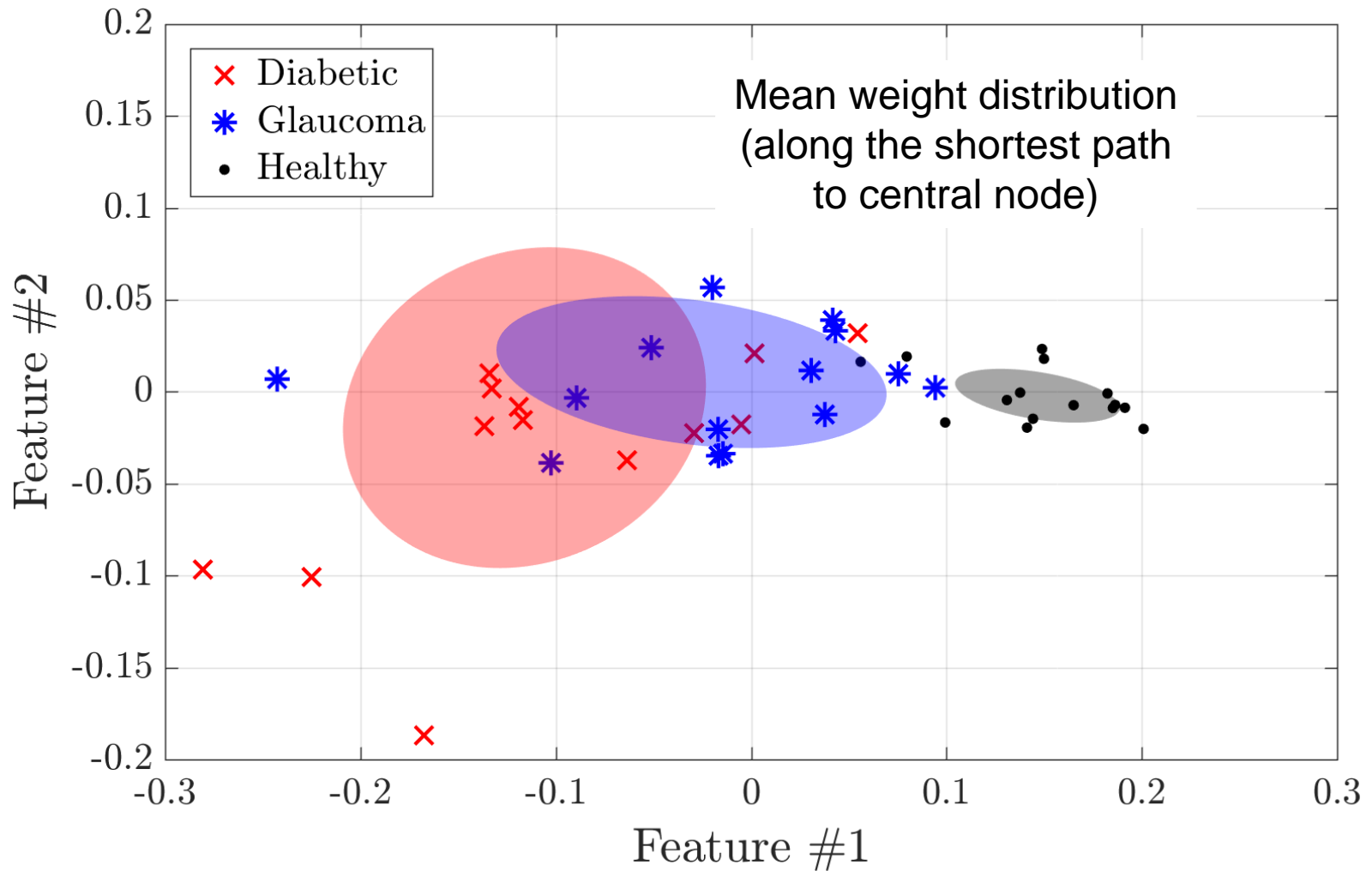


# Performance of network features in manual segmentation:



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

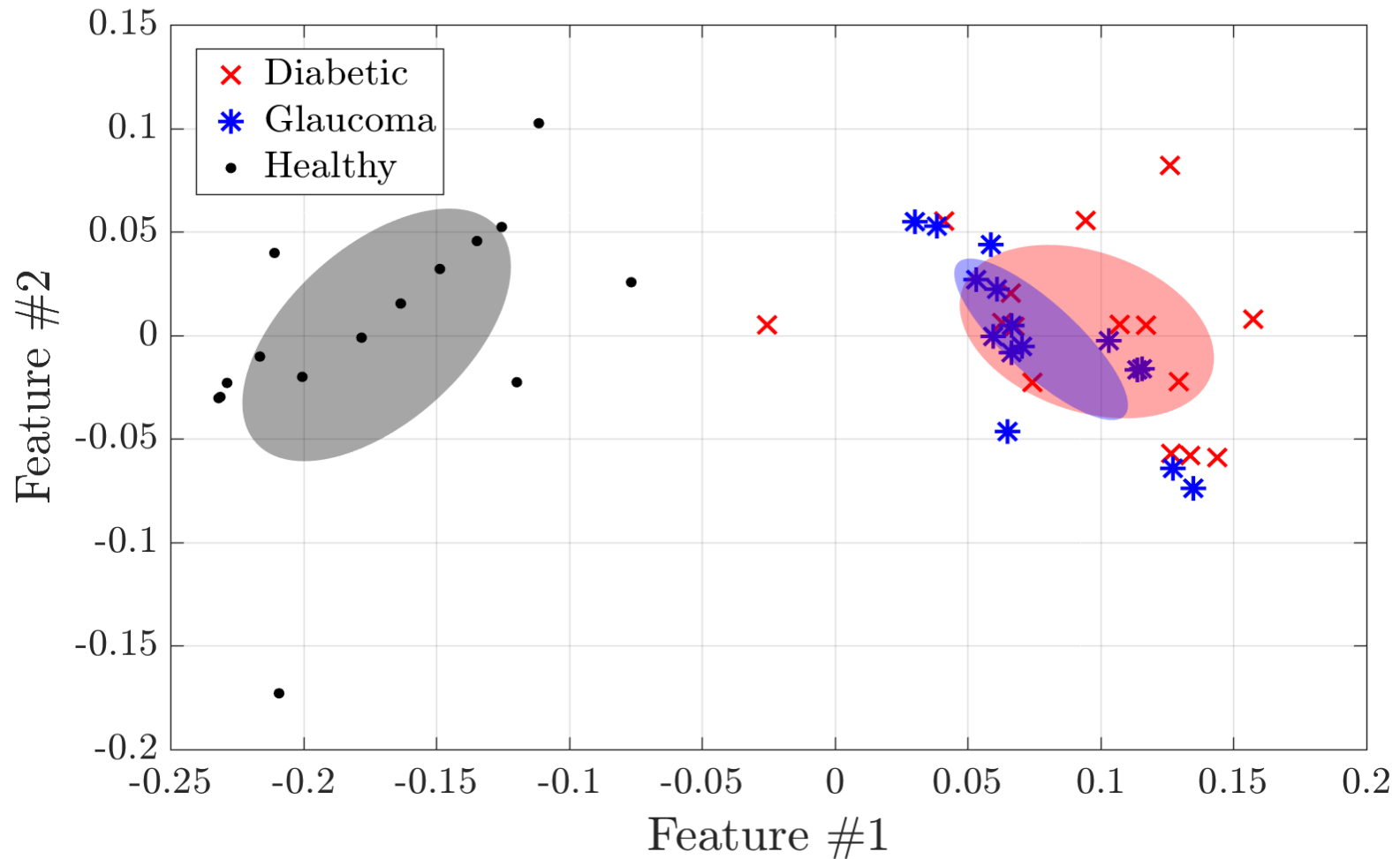
# Performance of network features in manual segmentation:



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

# Manual segmentation: distribution of weighted degrees

$$s_i = \sum_j w_{i,j}$$

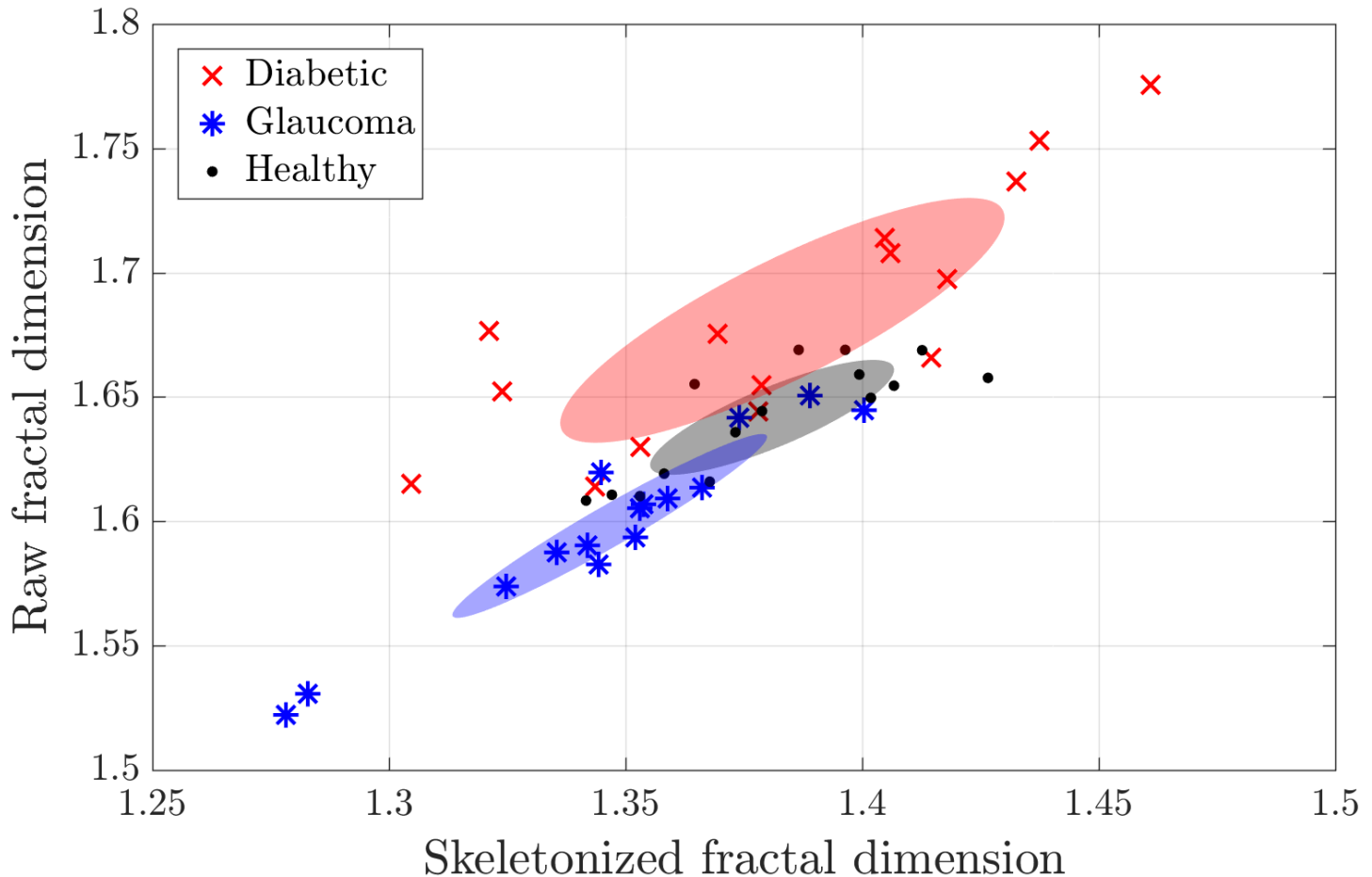


*P. Amil et al, Network-based features for retinal fundus vessel structure analysis, PLoS ONE 14 e0220132 (2019).*

# In the automated segmentation:

Fractal dimension analysis separates the three groups

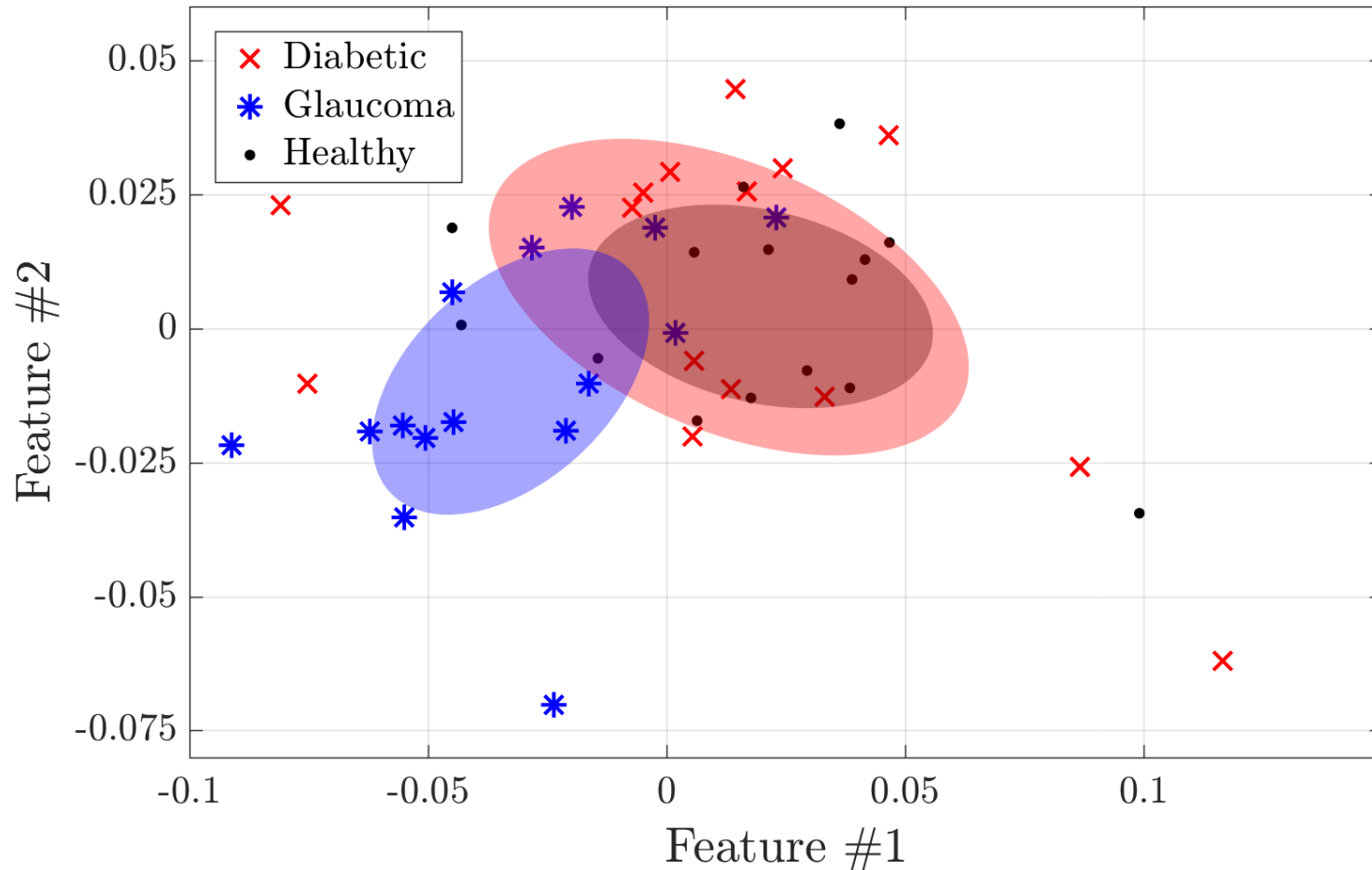
$$D = \lim_{\varepsilon \rightarrow 0} \frac{\log(N(\varepsilon))}{\log(1/\varepsilon)}$$



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

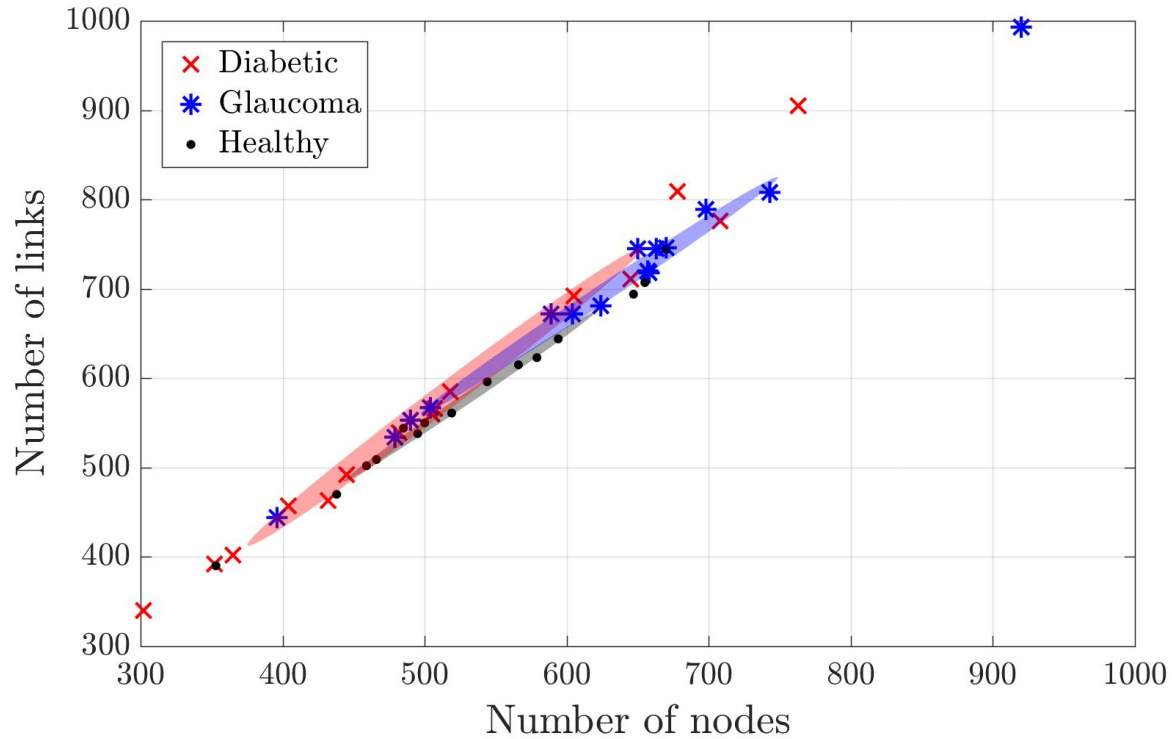
# In the automated segmentation:

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



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

# Simple network features do not differentiate



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

A measure for comparing graphs (unlabeled nodes, undirected links) was used for

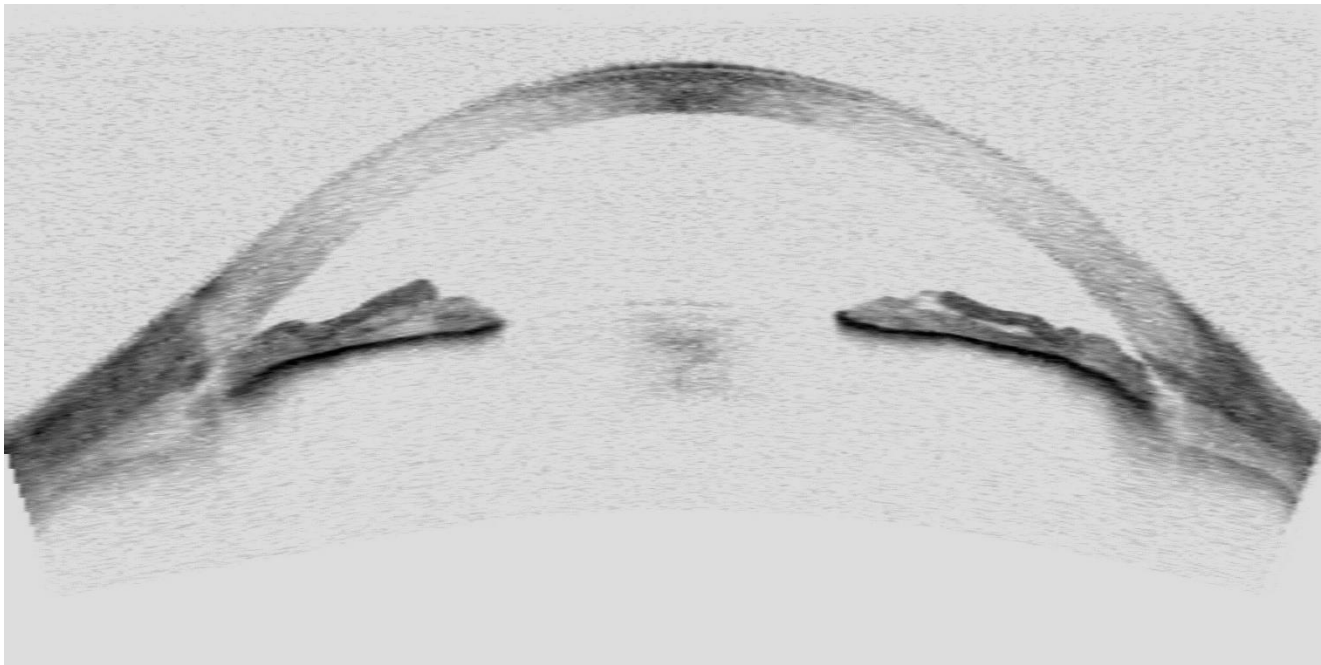
- EEG classification: two brain regions were identified that have different connection strength in control and in alcoholic subjects.
- Retina fundus image classification: perfect classification of healthy and non-healthy patients was obtained using high resolution images and manual segmentation.

*T. A. Schieber et al, Nat. Comm. 8, 13928 (2017)*

*P. Amil et al, PLoS ONE 14 e0220132 (2019)*

# 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).

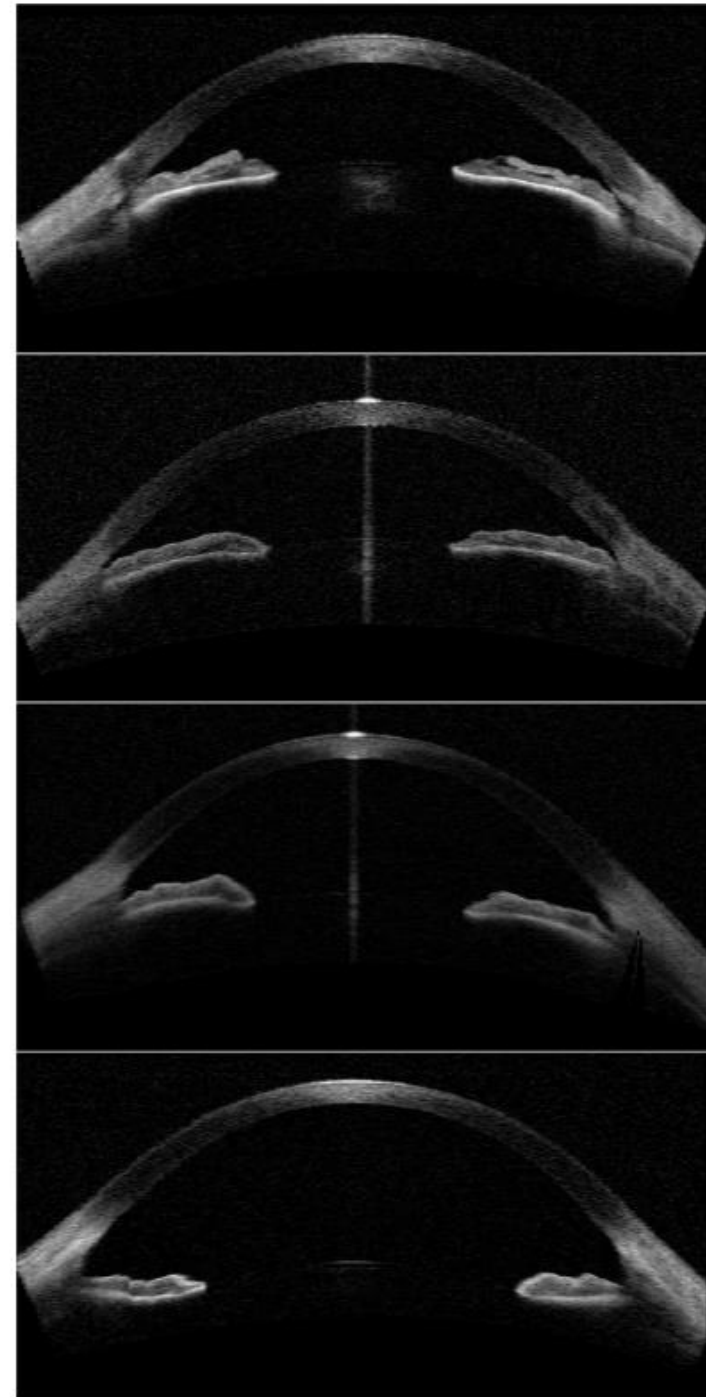




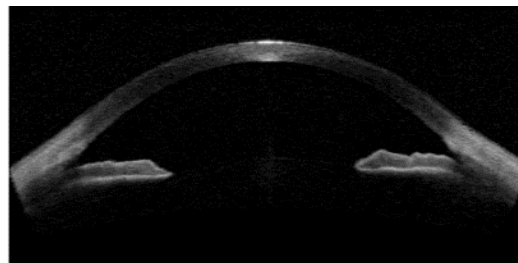
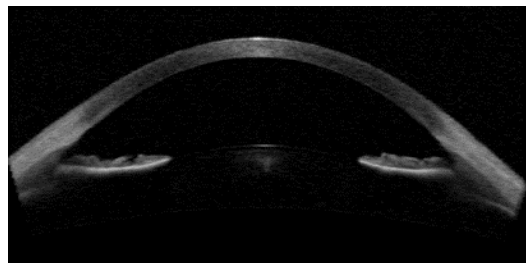
# Classified into 4 categories by two ophthalmologists and a trained PhD student

- Closed
- Narrow
- Open
- Wide open

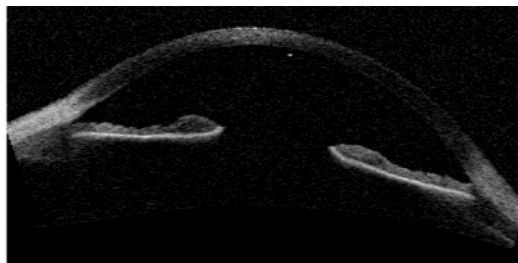
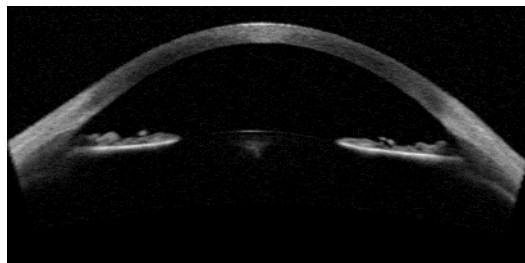
P. Amil et al., “*Unsupervised feature extraction of anterior chamber OCT images for ordering and classification*”, Sci. Rep. 9, 1157 (2019).



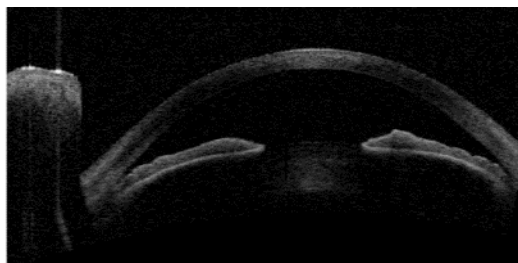
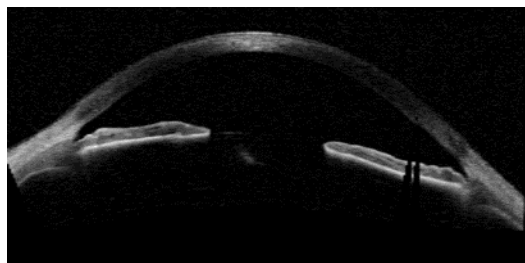
Using an appropriated distance (aligned Hellinger), by comparing pairs of images we extract features that can be used to order the images in a plane.



Small distance



Medium distance

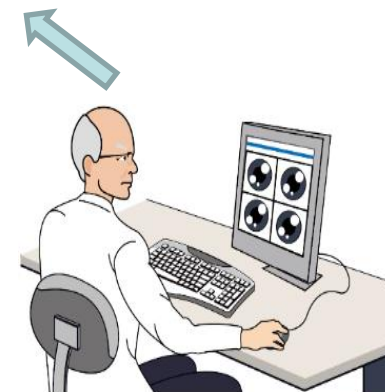
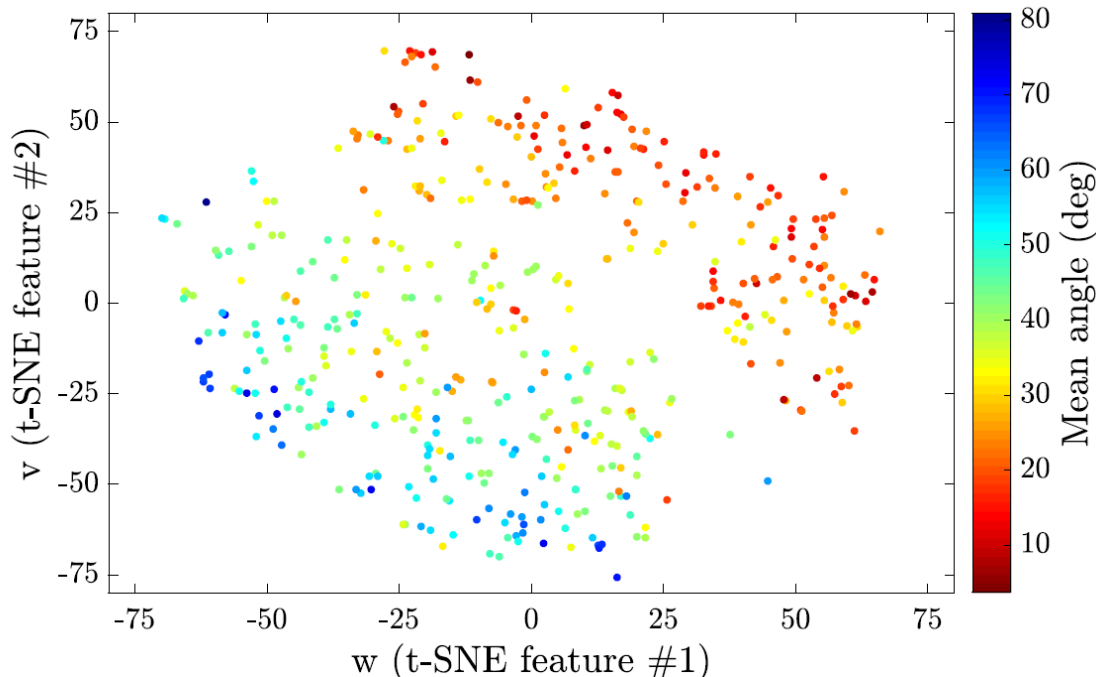
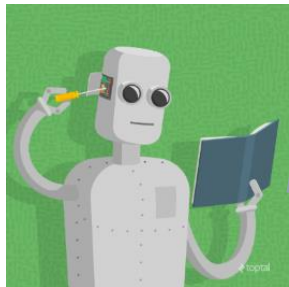
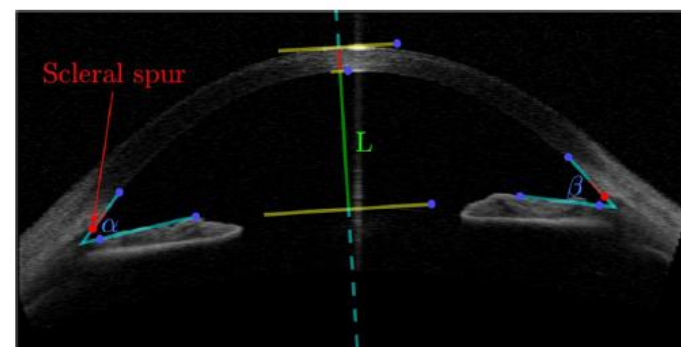


Large distance



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.



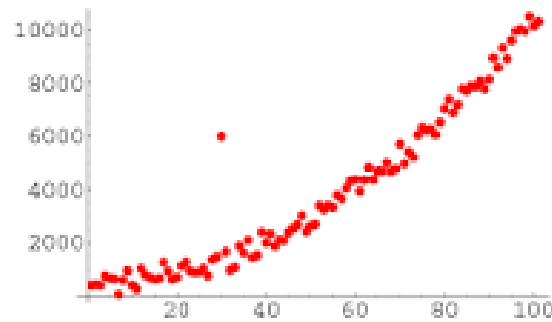
Can we do *better*?

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

# What is an outlier?



shutterstock.com • 1012593129



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

**Two “network-based” & “distance based”  
outlier detection methods**

- network percolation (fragmentation)**
- manifold learning**

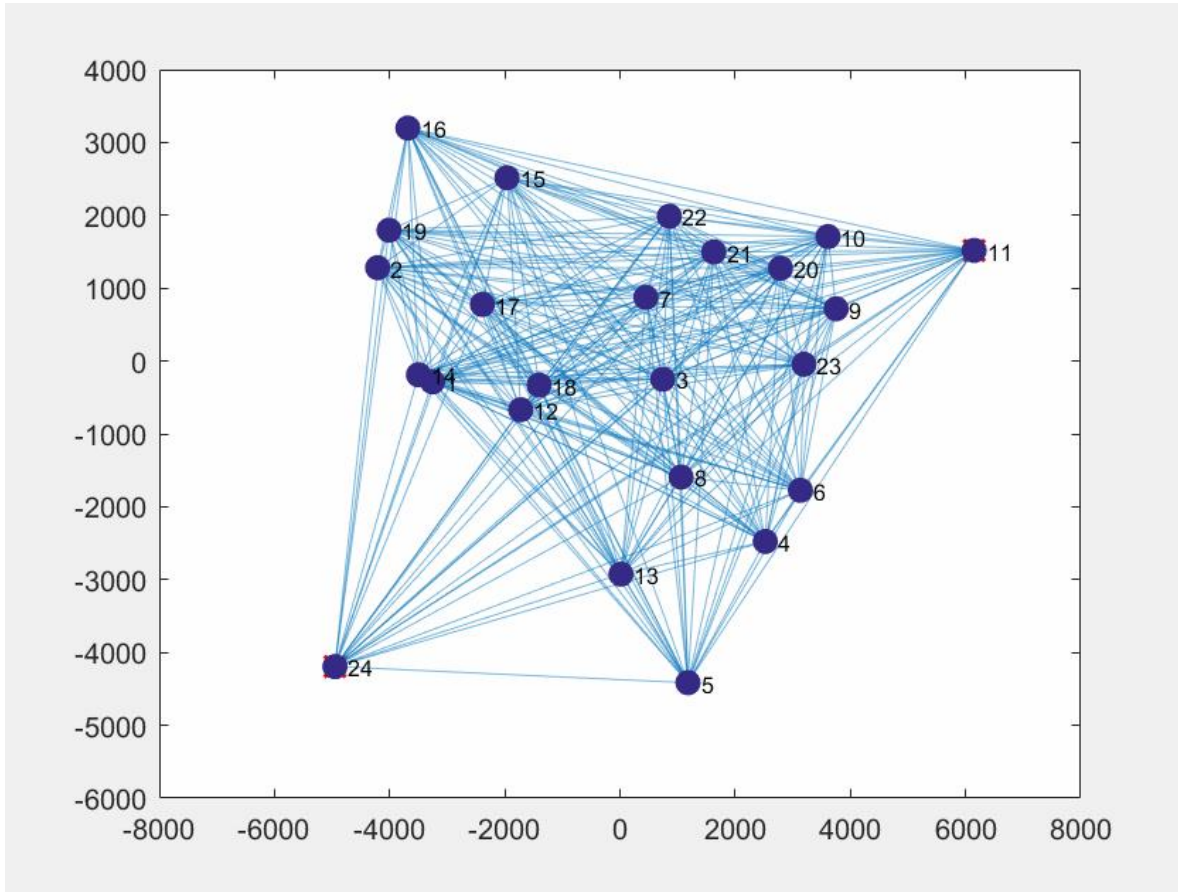
Vector of features that describe an item of a dataset

$$V_i = \{v_1^i \dots v_m^i\}$$

Distance between any two items

$$D_{ij} = \left( \sum_k |v_k^i - v_k^j|^p \right)^{1/p}$$

# First method: Outlier detection using percolation



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

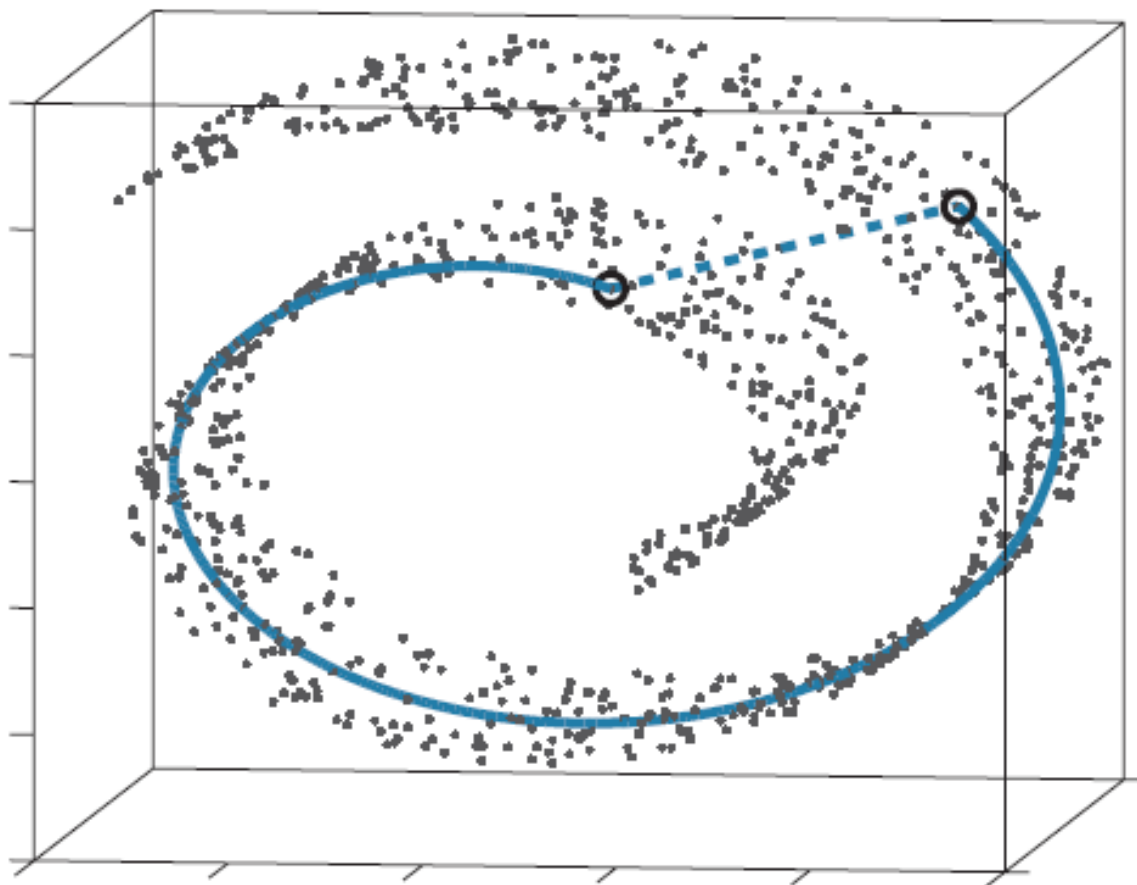
**Parameter free.**

P. Amil et al., “*Outlier mining methods based on network structure analysis*”, *Front. Phys.* 7, 194 (2019).



## Second method: manifold learning

- *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).

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

# 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  $\mathbf{D}'_{ij}$
- $AL_i = 1 - \rho_i^2$
- Two parameters (integers):
  - Dimension of reduced space
  - # of geodesic neighbors

Note: we *do not* detect outliers by analyzing the features returned by Isomap

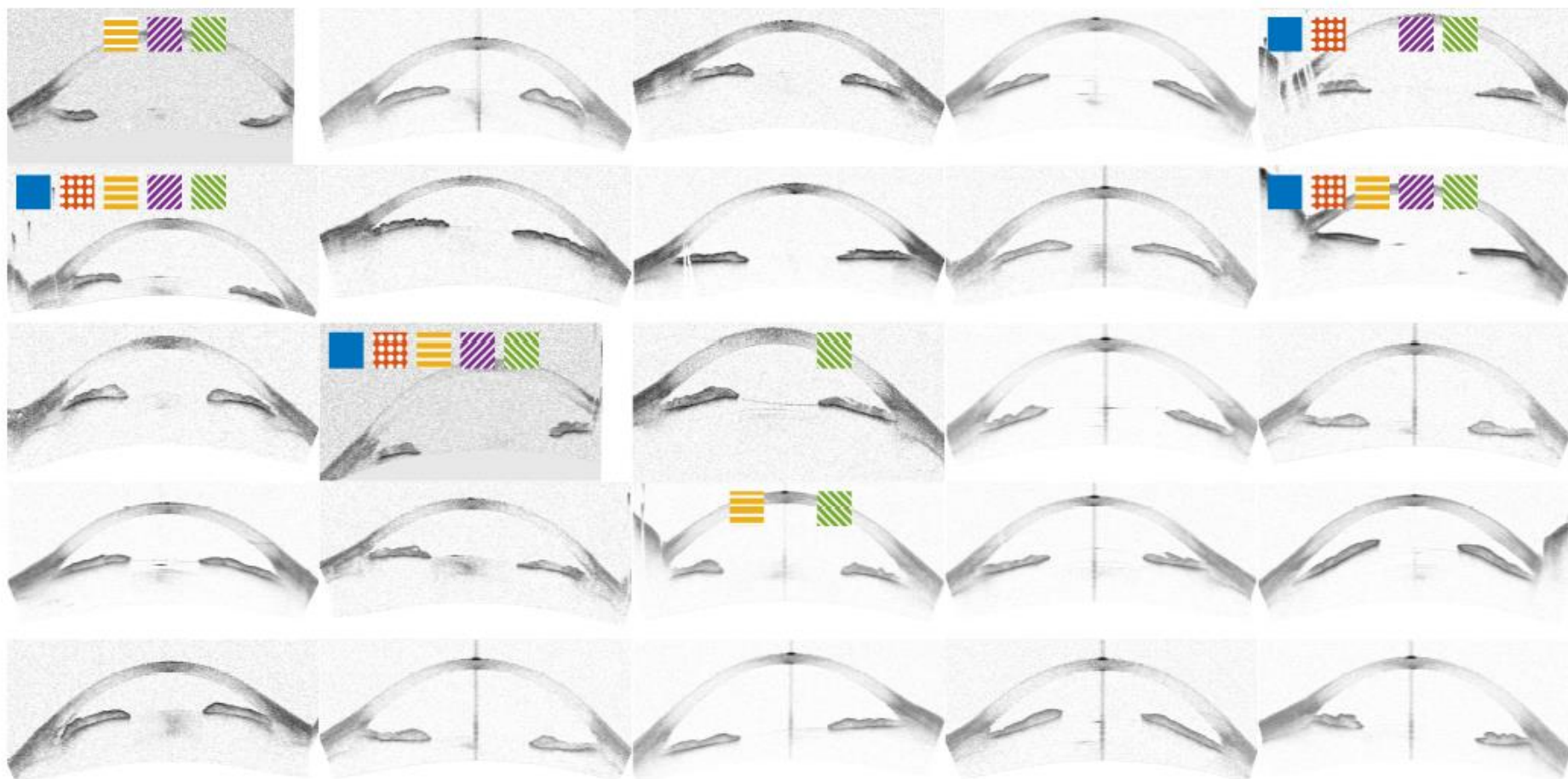
## Comparison with other outlier detection methods

- 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 distance-based method that 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 computation 13, 1443-1471, 2001.*

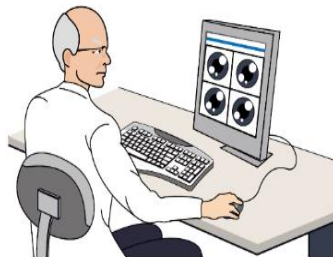
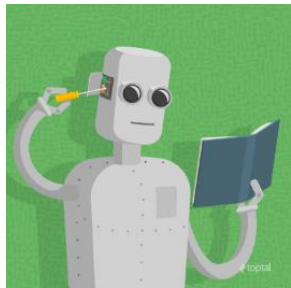
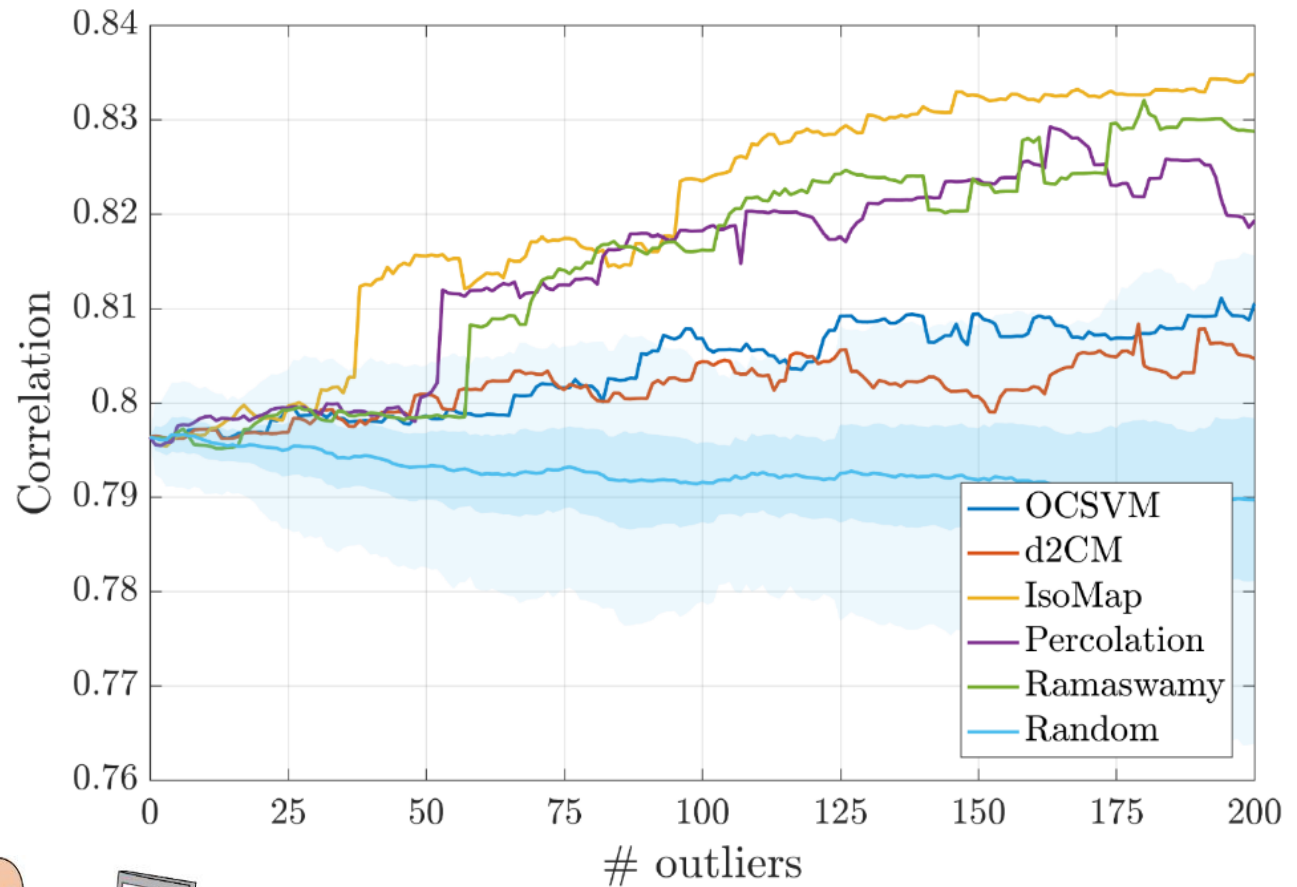
## Results with OCT images (1/2)



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)

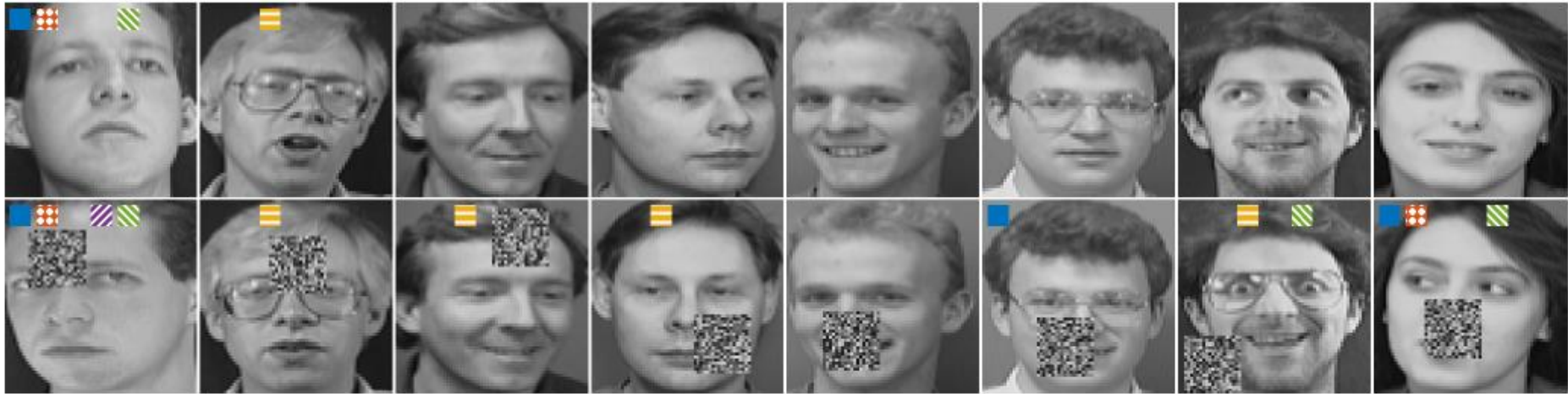
# Results with OCT images (2/2)

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

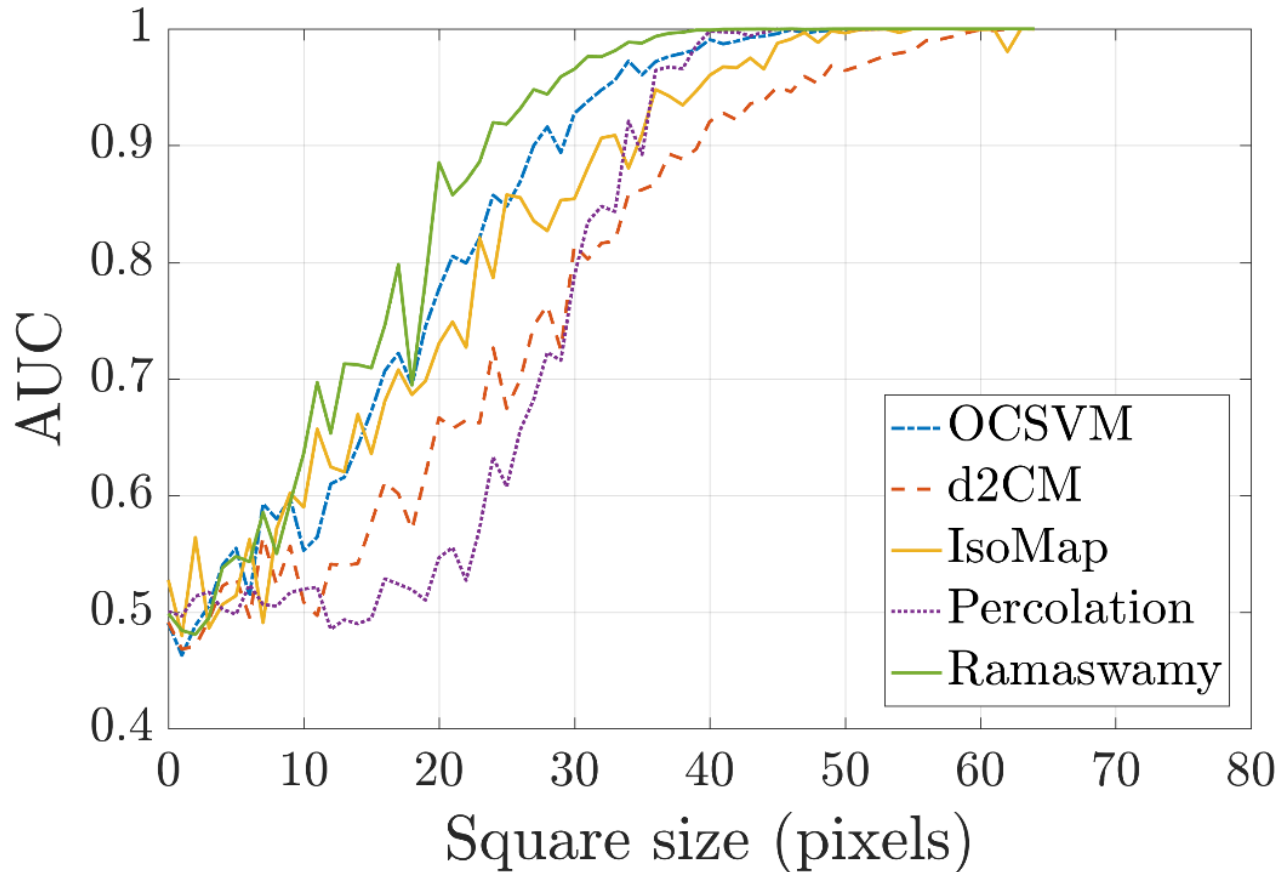


# Can these methods work with other images?

- 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.



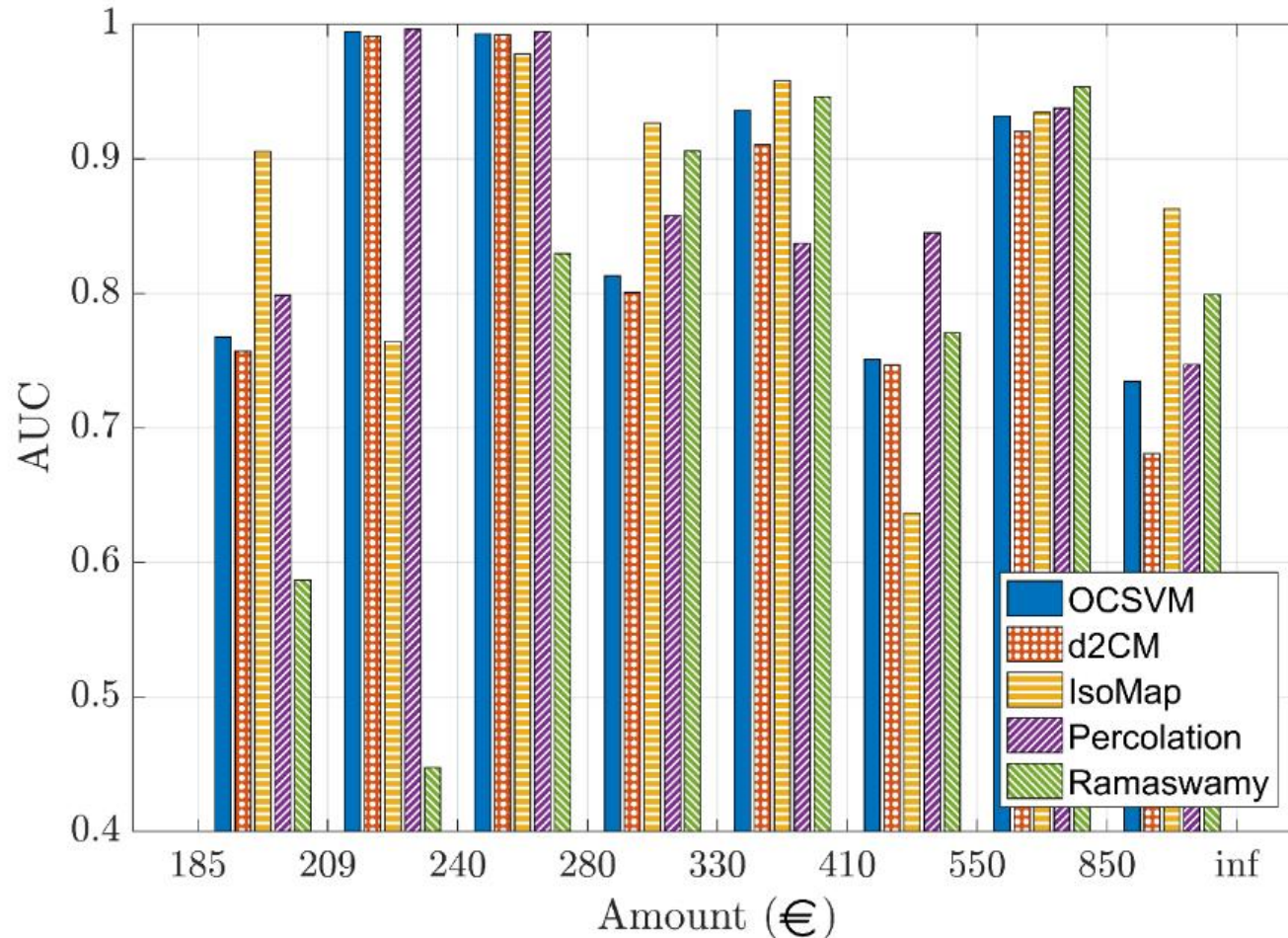
# Results from the face database



(similar results obtained with the area under the “precision-recall” curve)

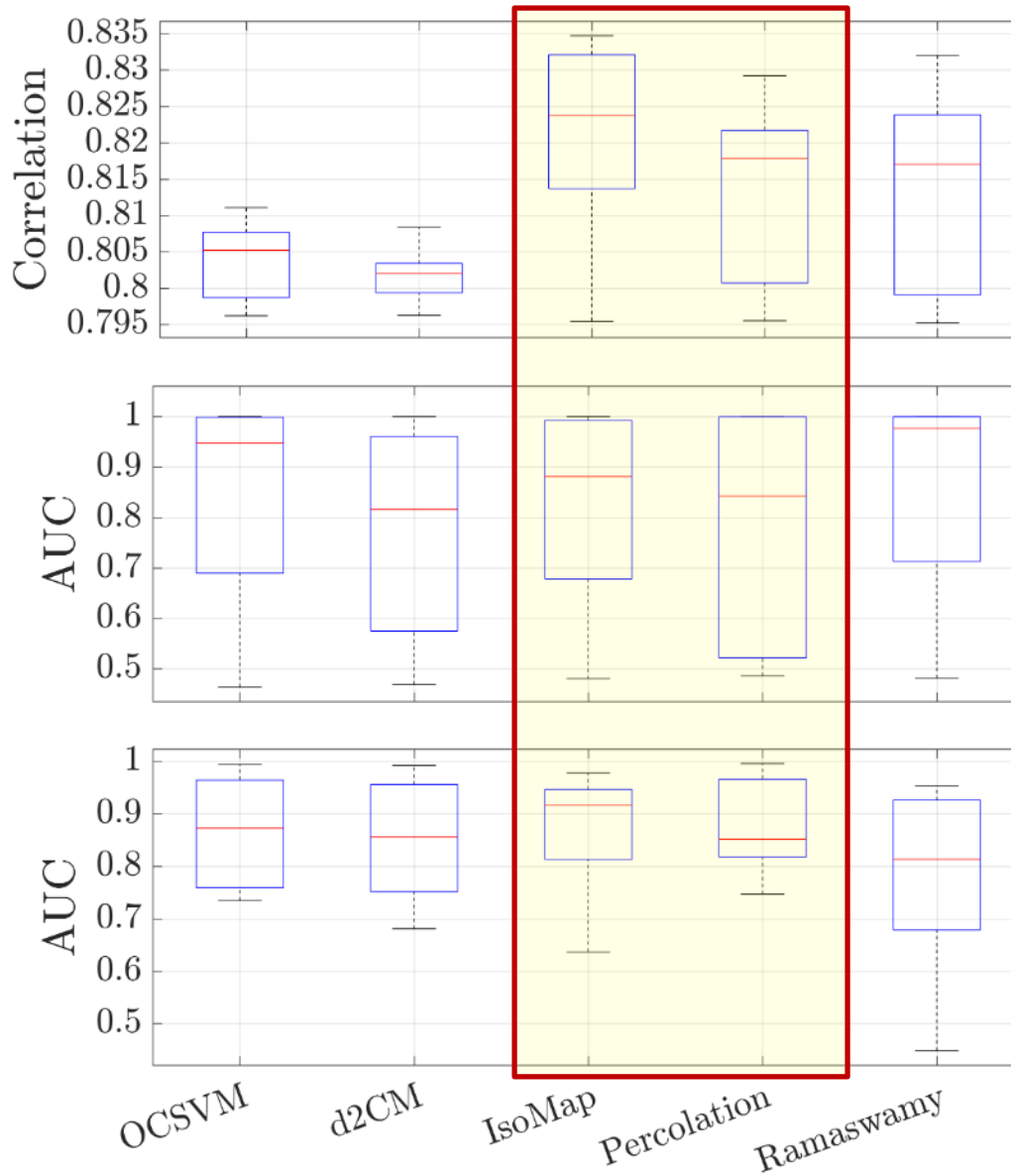
How about other types of elements?

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





# Summary of results



OCT database

Face database

Credit card database

## Summary II

- The performance of the outlier detection methods depends on the data; it is in general competitive with other methods used in the literature.
- 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 outperform other methods, but the performance is very sensitive to the parameters.
- Both methods are suitable for high dimensional, not too large databases (execution time grows linearly with the dimension of the data and at least as  $N \times N$  with the size of the database).

# Thanks to

- Laura González, Elena Arrondo, Cecilia Salinas, Jose Luis Guell (IMO, Barcelona)
- Ulrich Parlitz (Max Plank Institute for Dynamics and Self-organization)
- Fabian Reyes, Lev Guzman-Vargas (Instituto Politecnico Nacional, Mexico)
- Irene Sendiña-Nadal (Universidad Politecnica de Madrid)
- Nahuel Almeida (Universidad Nacional de Córdoba, Argentina)

**Thank you for your attention !**  
**Cristina.masoller@upc.edu**

- 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 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).

