Analyzing formal features of archaeological artefacts through the application of Spectral Clustering

Diego Jiménez-Badillo

diego_jimenez@inah.gob.mx
National Institute of Anthropology and History
México City

Salvador Ruíz Correa

src@cimat.mx
Centro de Investigaciones en Matemáticas
Guanajuato, Mexico

Omar Méndoza-Montoya

omendoz@cimat.mx
Centro de Investigaciones en Matemáticas
Guanajuato, Mexico

Introduction

- This paper is part of a broader effort to introduce the archaeological community to a range of computer tools and mathematical algorithms for analyzing archaeological collections. These include:

1. Application of clustering techniques for unsupervised learning.
2. Acquisition and analysis of 3D digital models.
3. Application of computer vision algorithms for automatic recognition of artefacts.
4. Automatic classification of shape features.

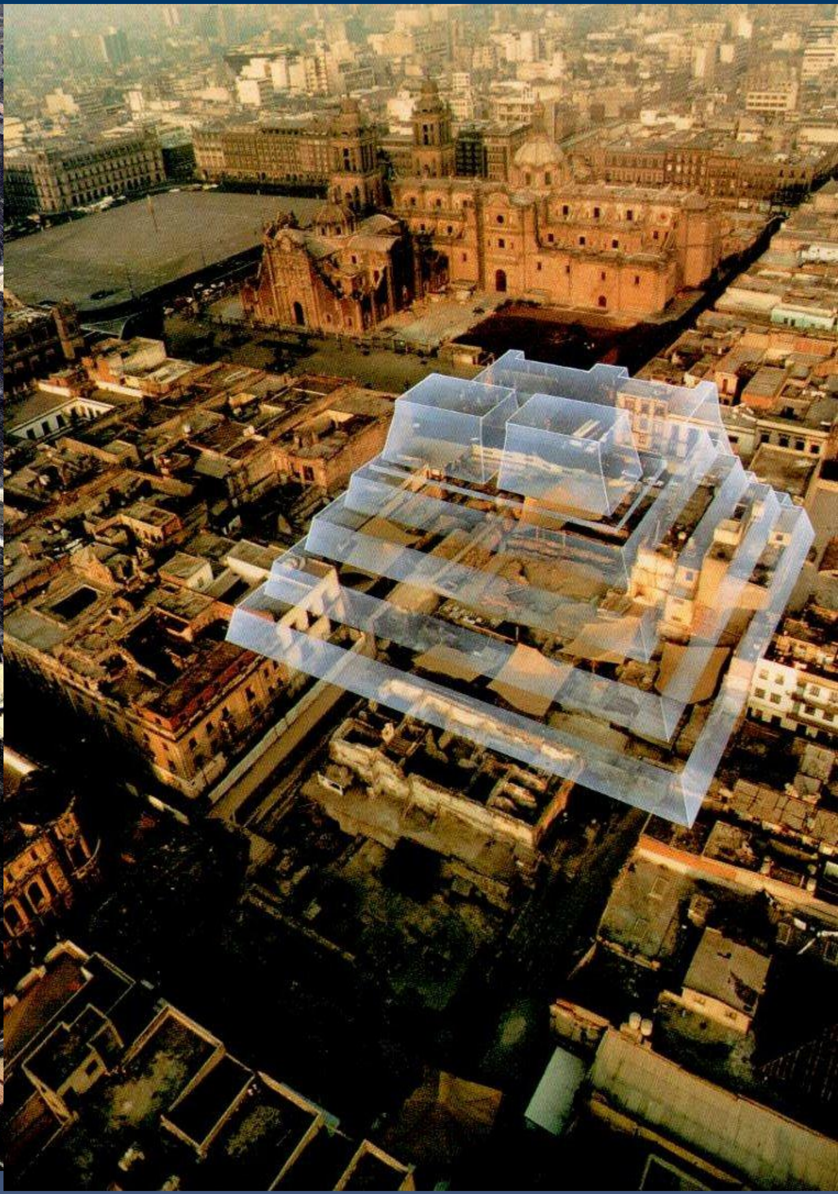
This presentation

- Here, we will focus on a new methodology for the analysis of archaeological masks based on a quantitative procedure called Spectral Clustering.
- This technique has not been applied before in archaeology despite its proven performance for partitioning a collection of artifacts into meaningful groups.

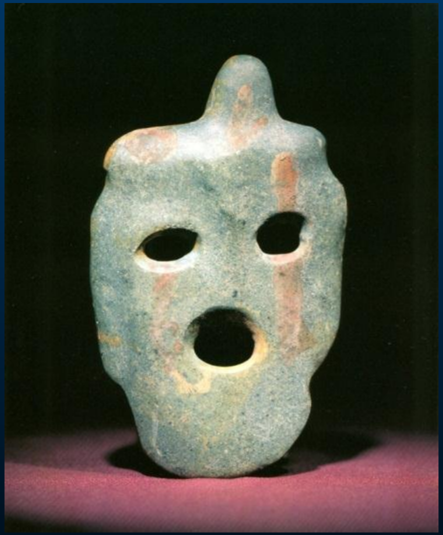
The Mask Collection from the Great Temple of Tenochtitlan

Study case

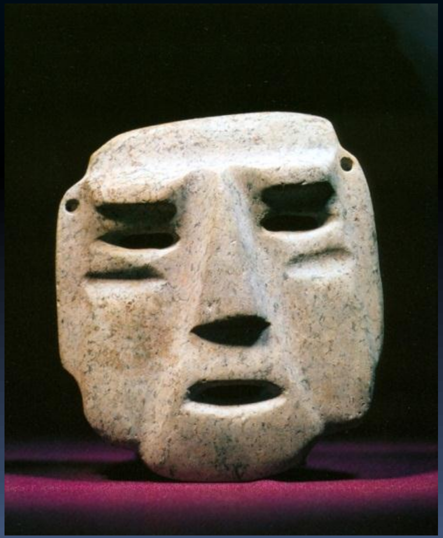
The idea for this project came from the need to classify similarities in 162 stone masks found in the remains of the Sacred Precinct of Tenochtitlan, the main ceremonial Aztec complex, located in Mexico City.



The schematic features of these objects set them apart from other more “naturalistic” style artifacts.



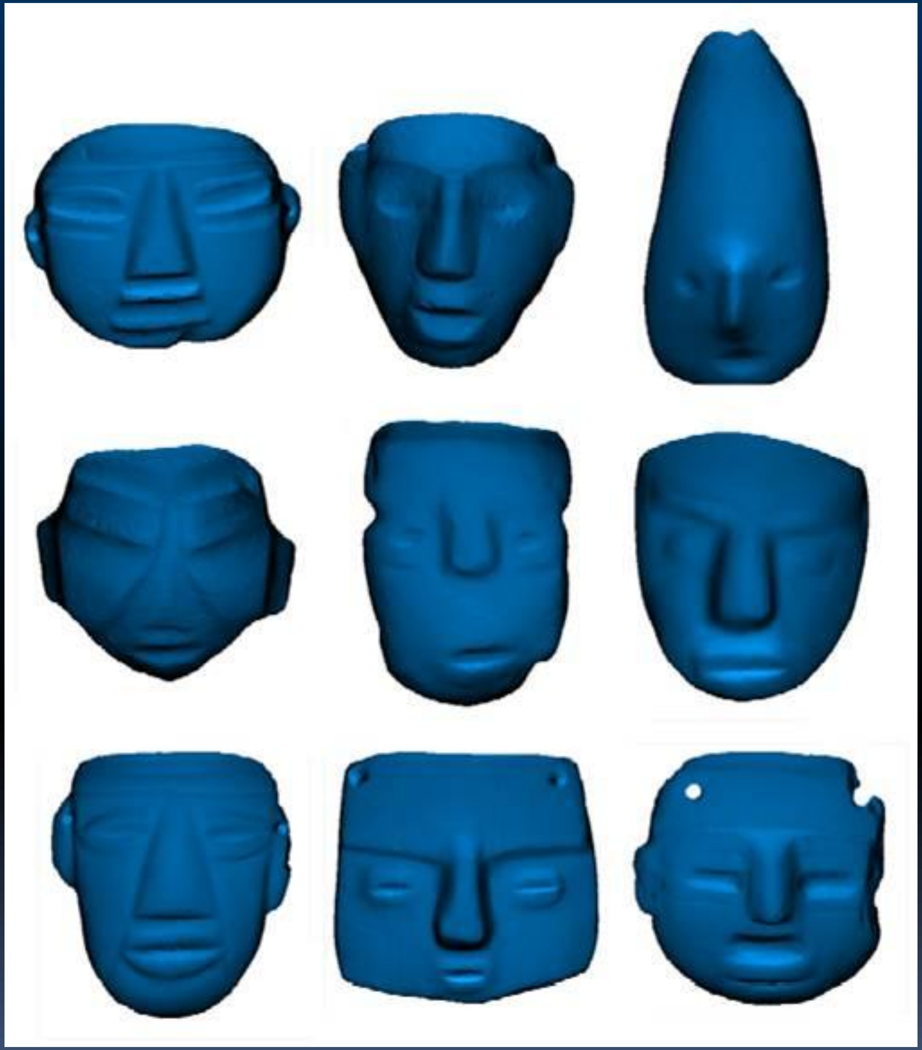
Their appearance has attracted the attention of many specialists and during the last three decades these masks have been the subject of intense debate for two main reasons:



These masks are interesting for several reasons:

- First, 220 figurines and 162 masks were located in 14 Aztec offerings dating from 1390 to 1469 A.C., yet they do not show typical “Aztec features”.

Indeed, their appearance resembles artefacts from Teotihuacan and from the southern State of Guerrero, particularly from the Mezcala region, which is hundreds of kilometers away from the ancient Tenochtitlan.



- Second, it is not clear how many Guerrero/Mezcala styles exist:

  Some specialists believe there are at least five[1] different traditions while others recognize only four[2], and another group of researchers sees only two[3] (Serra Puche 1975).

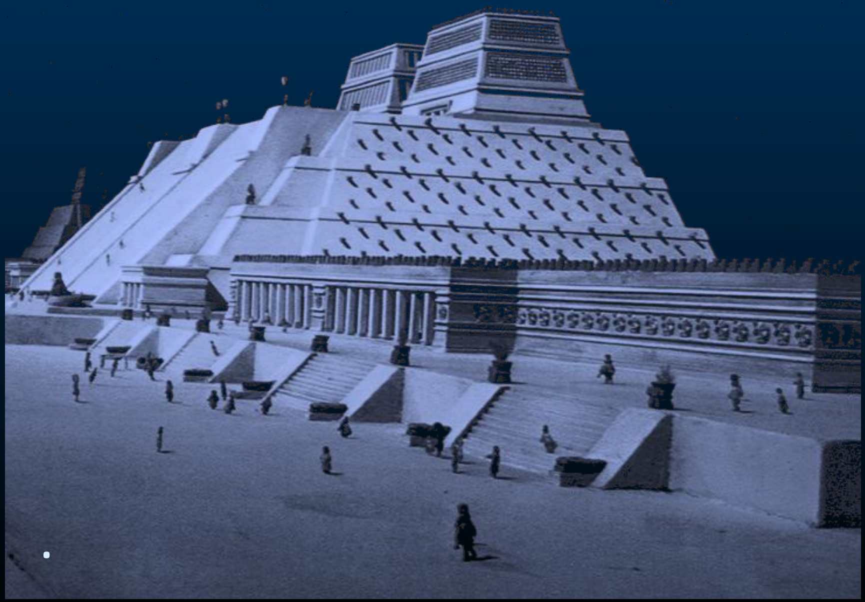
[1]Covarrubias 1948, 1961; Olmedo and Gonzalez 1986; Gonzalez and Olmedo 1990

[2] Gay, 1967

[3] Serra Puche 1975

More objective methods are needed to answer questions such as:

- How many styles were developed in the Guerrero/Mezcala regions?
- How many of these styles coexisted?
- Were some styles contemporary with the Aztecs?
- Were some of these masks manufactured by the Aztecs?
- Which specific styles are represented among the 162 masks found in the Sacred Precinct of Tenochtitlan?



Clustering basics

The application of clustering in archaeology

One of the most important applications of clustering in archaeology is “unsupervised learning”, that is the discovery of artifact groups based exclusively on the analysis of characteristics observed in the artifacts themselves.

Once the collection has been segmented into several groups, it would become easier to distinguish potential classes.

Clustering basics

Given a set of data-points, any clustering algorithm seeks to separate those points into a finite quantity of groups (i.e., clusters). It applies an objective similarity function to weigh how close (similar) or distant (dissimilar) the original data are among themselves.

Items assigned to the same group must be highly similar. At the same time, items belonging to different groups must be highly dissimilar.

The quality of the algorithm is judged by how successful it is to accomplish: a) greater homogeneity within a group, and b) greater heterogeneity between groups.

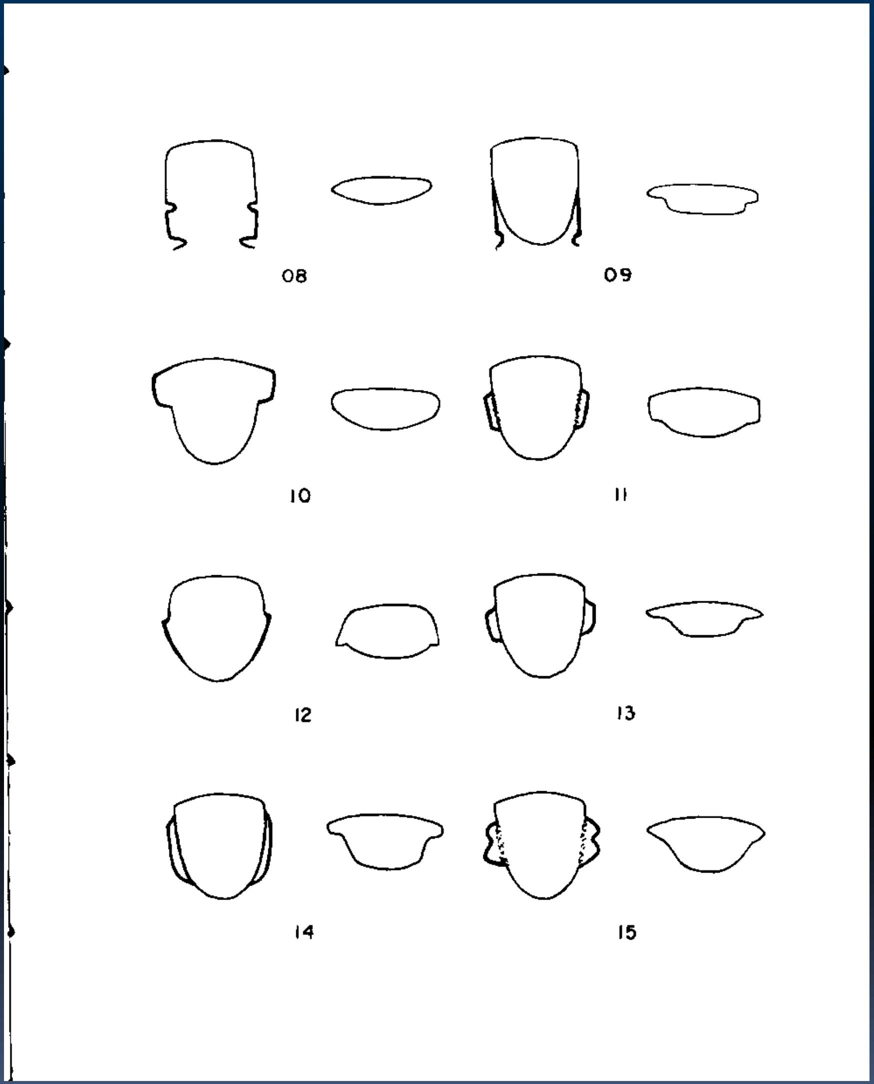
Two clustering approaches are very popular:

1. Component linkage algorithms (i.e. single and total linkage) are based on thresholding pairwise distances and are best suited for discovering complex elongated clusters, but are very sensitive to noise in the data.

2. K-means algorithms, on the other hand, are very robust to noise but are best suited for rounded linearly separable clusters.

Boundary shapes

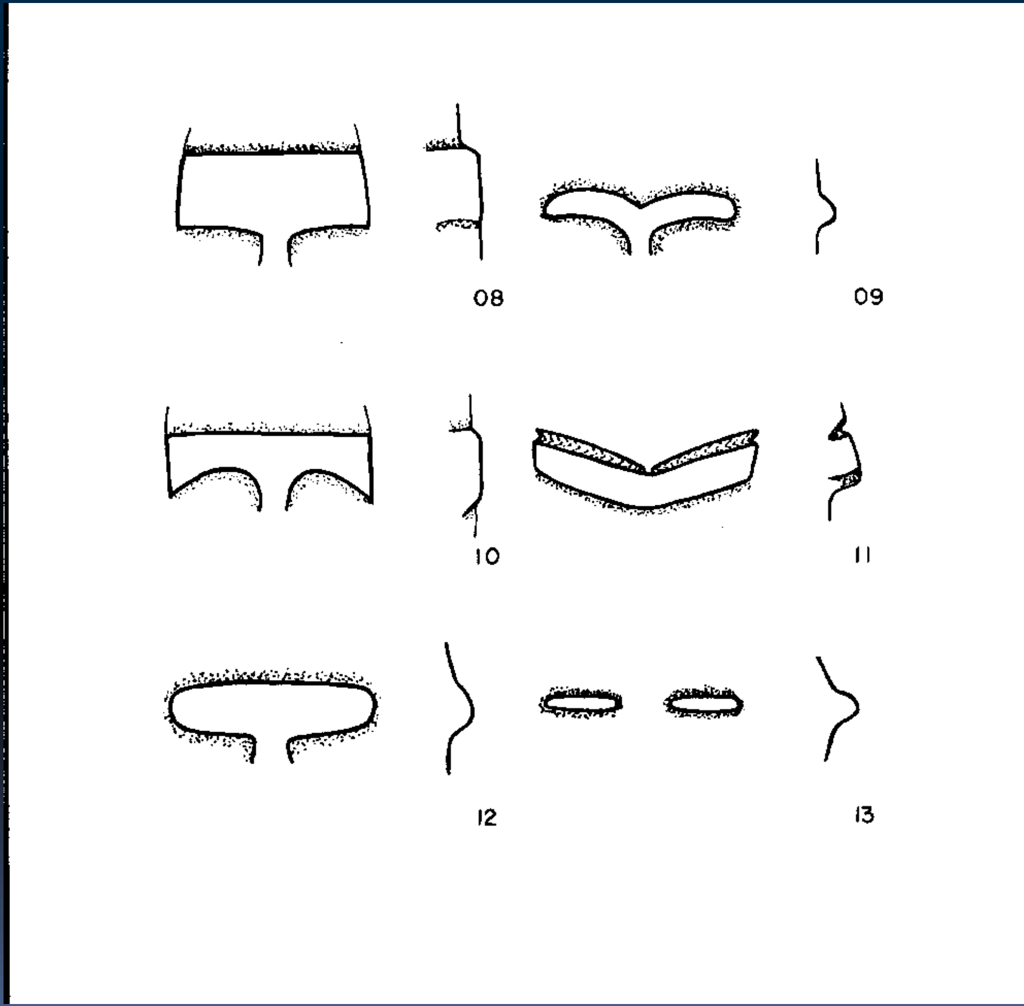
- Some years ago, Olmedo and Gonzalez (1986) proposed a classification based on a component-linkage algorithm (numerical taxonomy).
- The forms of faces, eyes, noses, eyebrows, etc. were codified categorically.
- This produced an input of 23 shape attributes for each one of 162 masks.

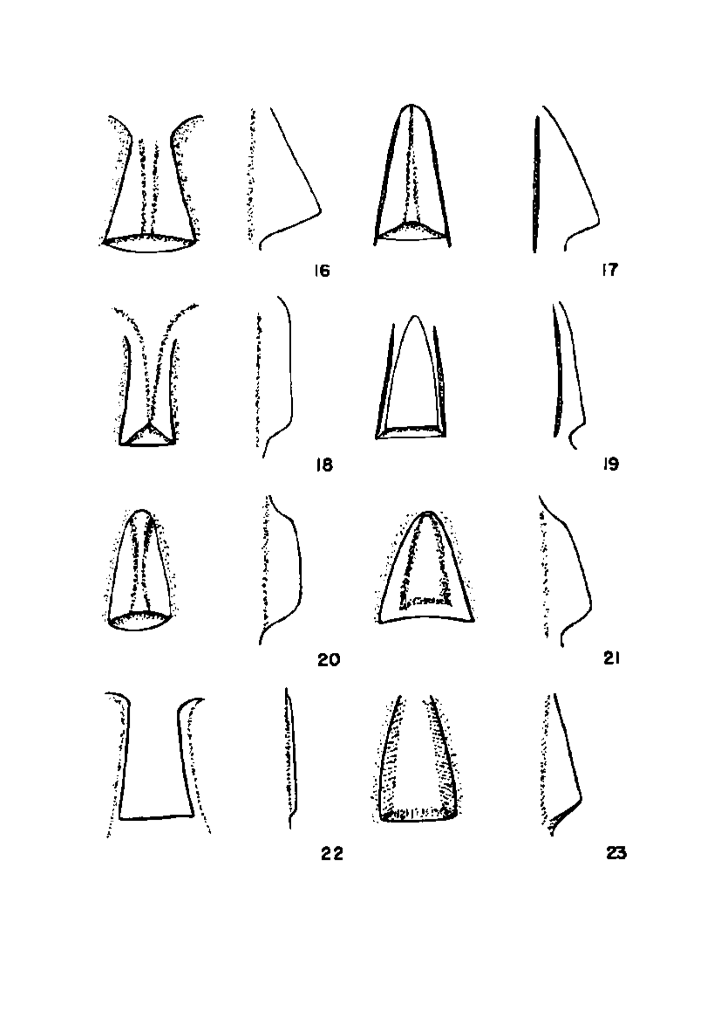


Olmedo and Gonzalez. 1986. Presencia del Estilo Mezcala en el Templo Mayor. INAH, México.

Eyebrows shapes

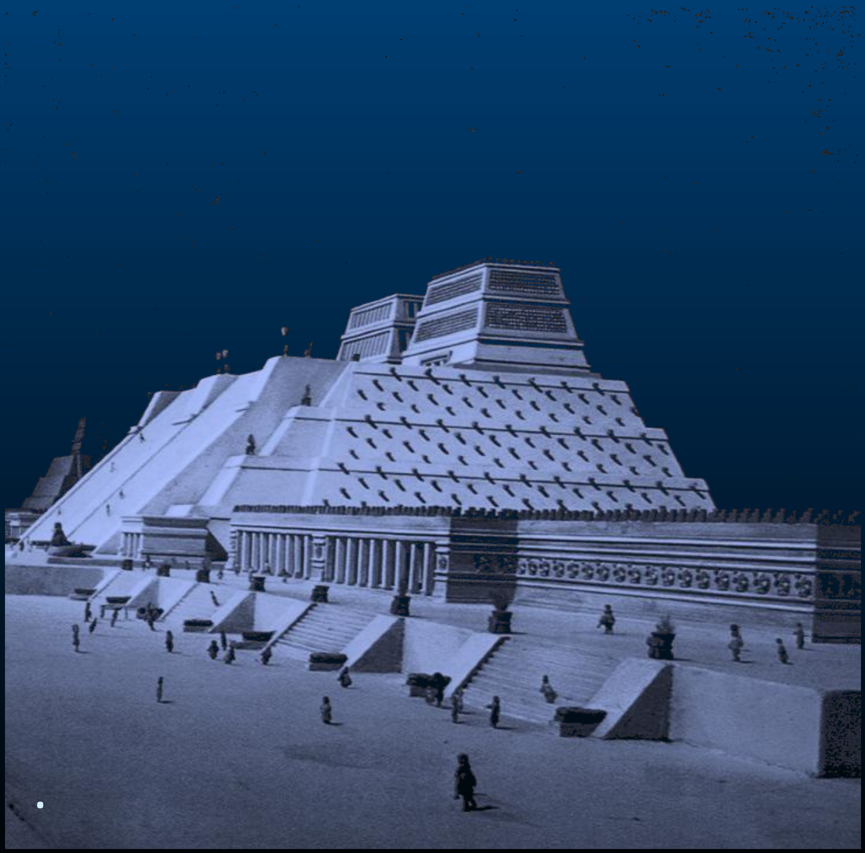
Noise shapes





Olmedo and González results

- This lead to the identification of 40 groups, of which:
  - 13 groups include only 2 masks
  - 13 groups include only 3 masks
  - 6 groups include 4 masks
  - 20 masks were not included in any group



Spectral Clustering

SPECTRAL CLUSTERING

Spectral Clustering is a state-of-the-art technique for exploratory analysis.

It is derived from Spectral Graph Theory, a branch of mathematics that studies the properties of graphs in relation to the eigenvalues and eigenvectors of an especial type of matrix called Laplacian.

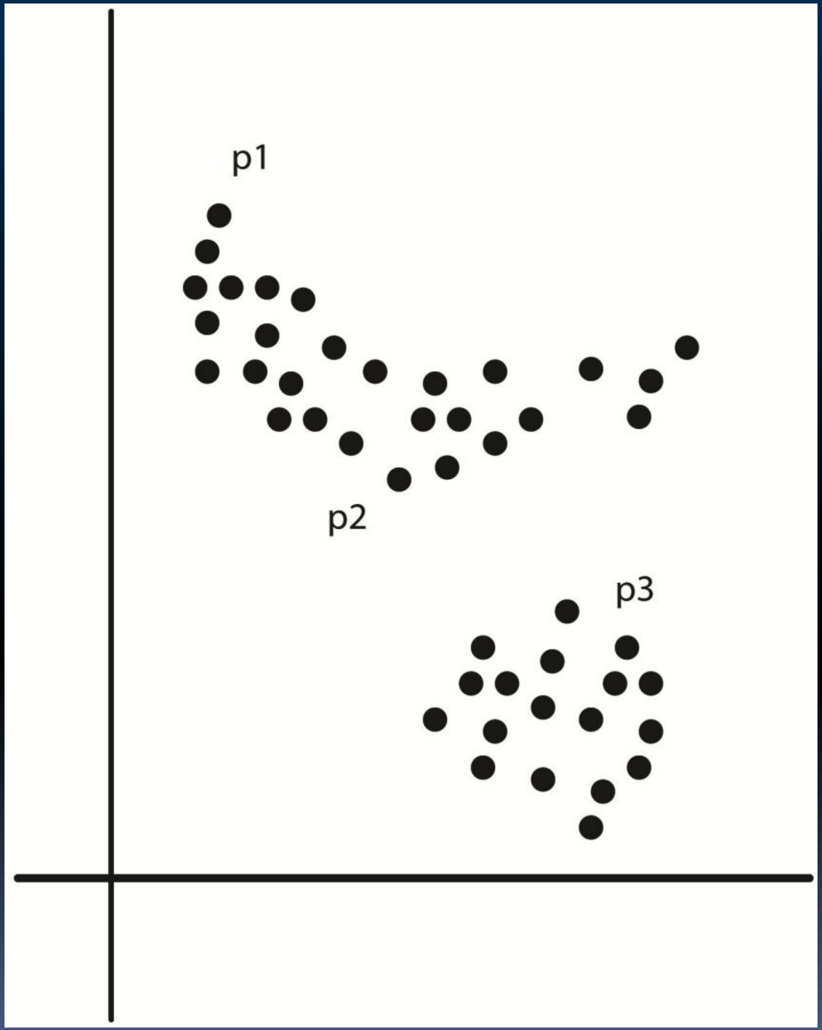
In contrast to other techniques, spectral clustering can be applied to different kinds of data, including categorical data. In many situations, categorical data is often the only source of information for archaeological unsupervised learning. Therefore we believe it could bring great benefits in our field.

Here, we won’t dive too deep into mathematical theory. Instead, we follow a very simple example to make the logic of this clustering technique easier to understand. Mathematical proofs can be found in the literature referenced in the paper, especially Luxburg (2007).

Spectral Clustering is more efficient than linkage and k-means algorithms. It finds elongated clusters and is more robust to noise than linkage algorithms.

Spectral clustering logic

Spectral clustering seeks to identify groups by analyzing no the exact location of the data points (like single, total linkage and k-means), but the connectedness between them.



- Spectral clustering relies on a graph representation of the data set.
- In this graph each mask is represented as a vertex. Then, we calculate a measure of similarity (affinity) between all the masks. This is done in two steps:
- First, calculate the so-called Hamming distance, which measures the percentage of non-shared attributes between two artefacts. Obviously, this measure is very useful to work with categorical data. The formula is:

$$Hamming\ distance = \left(\frac{1}{n}\sum_{n=1}^{n} I(a_i \neq b_i)\right)^2$$

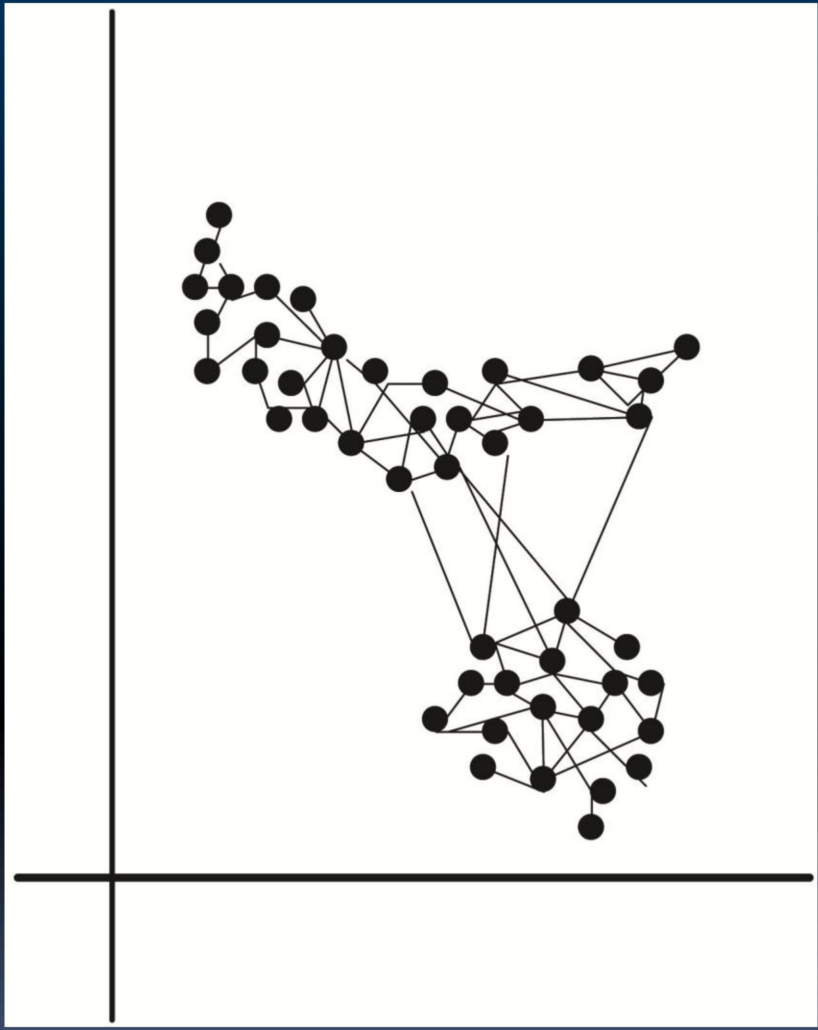
- Second, calculate the affinity between all the masks by applying the Gaussian function to the Hamming distance:

$$S(A,B) = exp\left\{-\frac{1}{2\sigma^2}\left(\frac{1}{n}\sum_{n=1}^{n} I(a_i \neq b_i)\right)^2\right\}$$

σ = Threshold to control the desired level of similarity

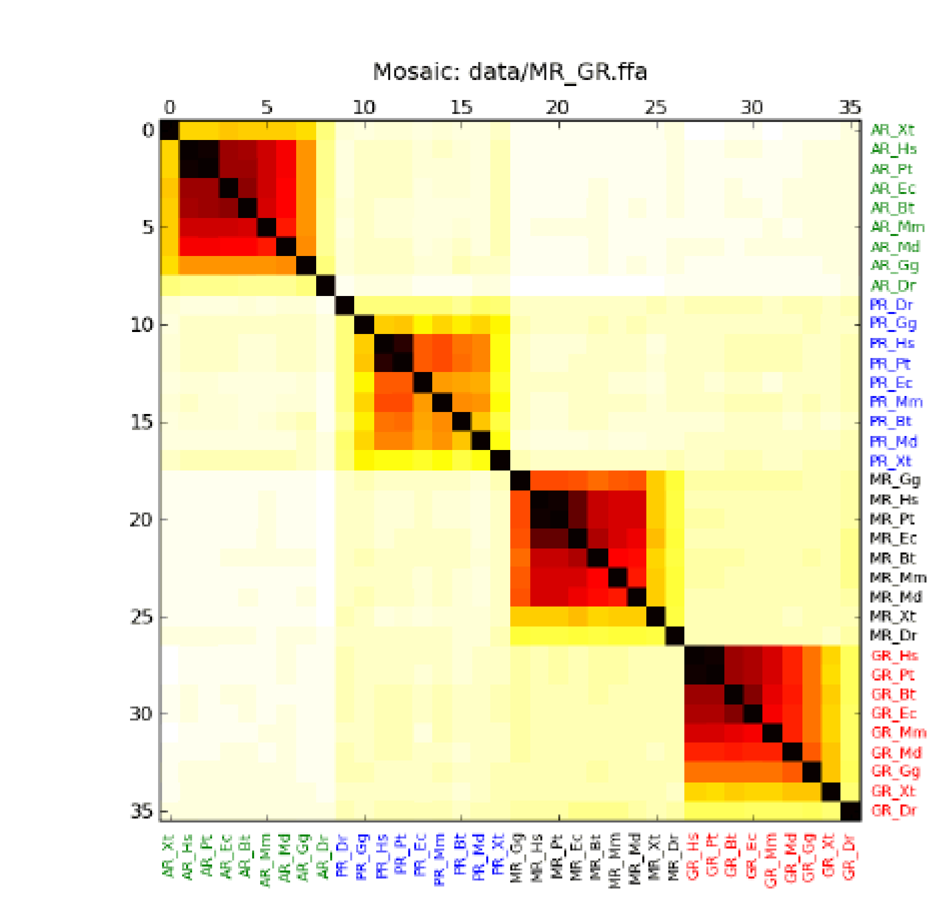
- If two objects are very different the result of the equation is negative and close to or equal to zero. On the contrary, if two objects are very similar the result of the equation is near or equal to 1.

- The next step is to draw a link between two vertices (i.e. masks) if the similarity between them is positive or larger than a certain threshold controlled by the parameter sigma.

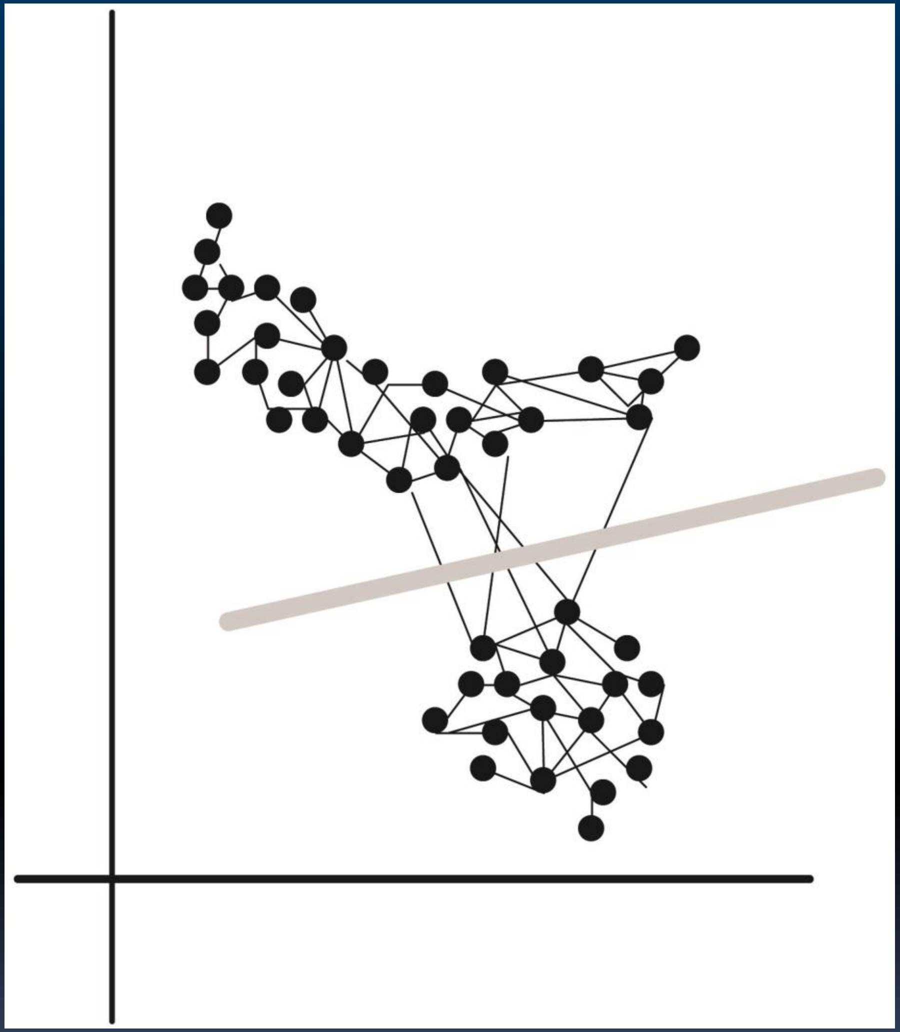


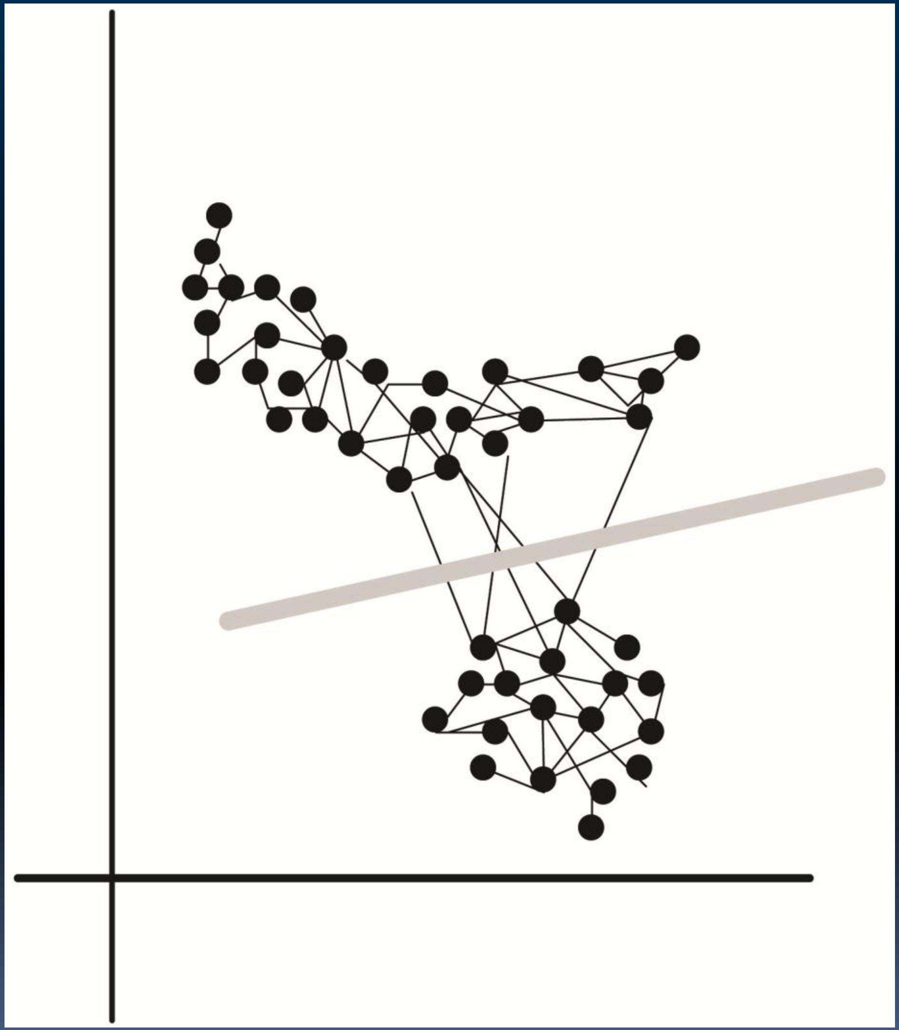
Next, we weight each edge of the graph with its corresponding similarity score.

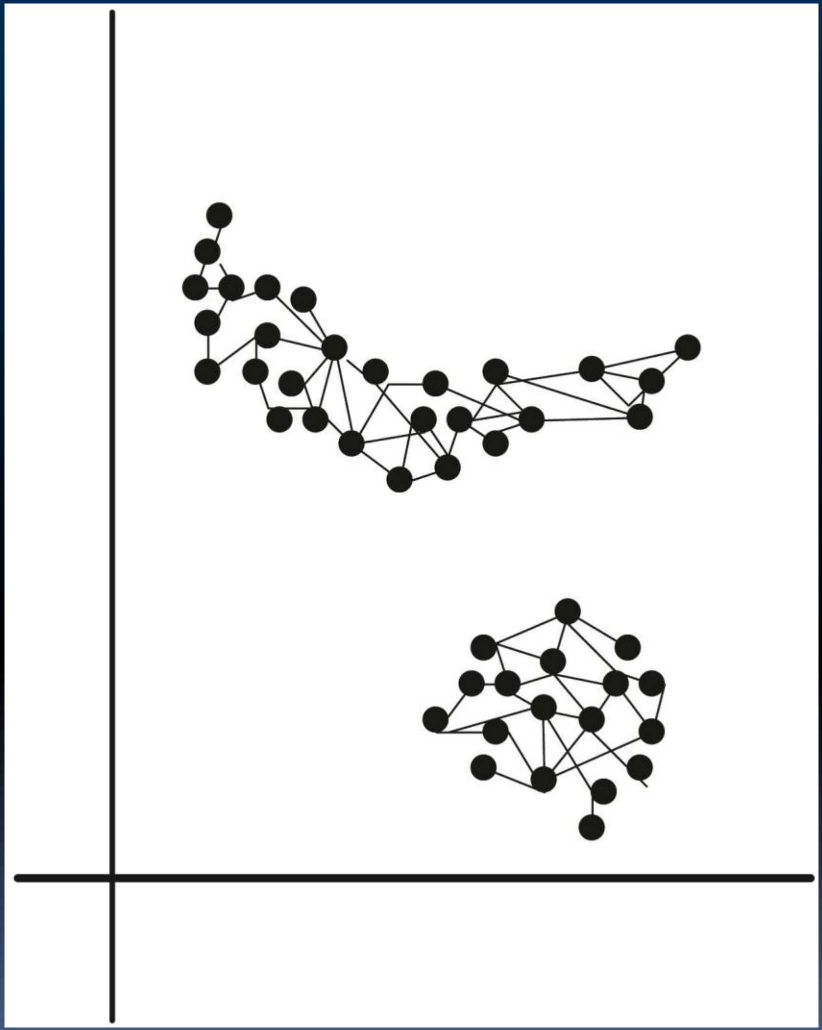
This step produces a matrix in which the affinity (i.e. similarity) of all objects with all others is recorded.



- Once this is done, the clustering problem can be reformulated as finding an optimal cutting of the graph. This means finding a partition of the graph such that the edges linking objects of the same group have very high weights (i.e. are quite similar), while the edges between groups have low weighs (i.e. are very dissimilar).







THE CUTTING PROBLEM

Finding an optimal cutting of the graph is what mathematicians call an NP-hard problem. This means that it cannot be found in real time. Therefore, we need to find an approximate solution.

One type of relaxation is based on analyzing the eigenestructure of the Laplacian matrix associated to the similarity graph. Mathematical proofs can be found in the extensive literature on the subject.

LAPLACIAN MATRIX

The so-called Laplacian matrix is a transformation of the similarity matrix that shows more clearly the structure of the dataset. To build a Laplacian Matrix we need two ingredients:

a. A matrix D containing information of the connectivity of the similarity graph. This is called the Degree matrix or Diagonal Matrix.

b. The similarity matrix S produced in step 2.

The simple Laplacian matrix satisfies the following equation:

$$L = D - S$$

EIGENSTRUCTURE

As you know “eigen” is a prefix that means “innate”, “own”, and “characteristic”. Thus, studying the eigenstructure of matrices serves the purpose of revealing the intrinsic nature of the data contained in those matrices.

Looking at the eigenstructure we can identify the best possible cuts for the graph and therefore the best clustering partition.

The eigenstructure is given by eigenvalues and eigenvectors.

EIGENVALUES AND EIGENVECTORS

Given a square symmetric matrix S, we say that λ (*lambda*) is an eigenvalue of S if there exists a non-zero vector x such that:

$$Sx = \lambda x \quad \text{equation (2).}$$

In equation (2), *x* represents the eigenvector associated to the eigenvalue λ and both constitute an eigenpair for the matrix S.

GRAPH SPECTRUM

Each eigenvector has its associated eigenvalue. So there are as many eigenvectors as eigenvalues. The spectrum of the graph is precisely the set of all eigenvalues that satisfy equation (2). Such property is invariant with respect to the orientation of the data set.

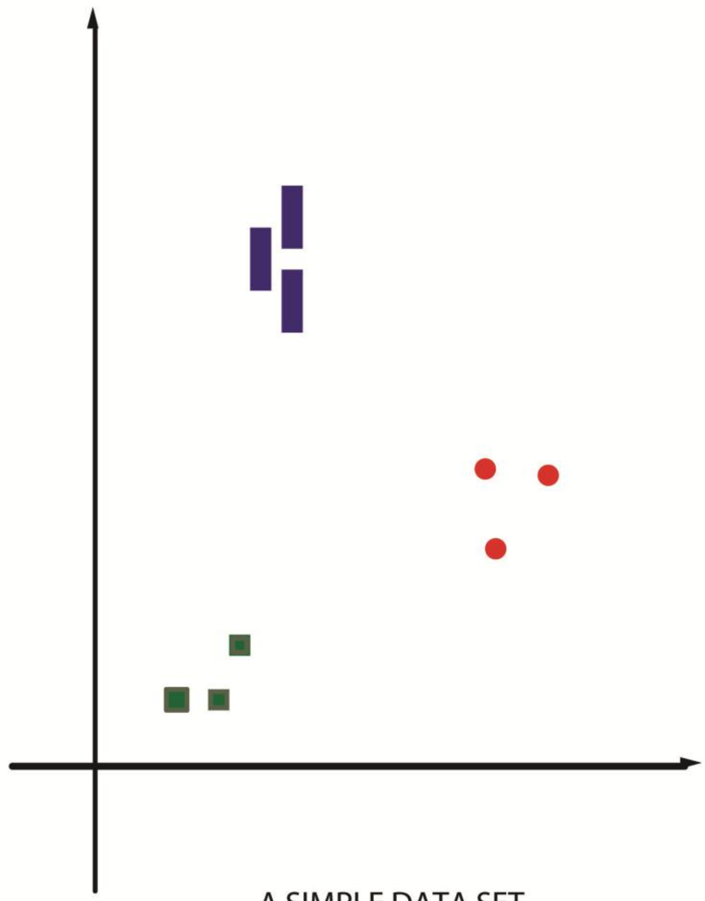
Calculating eigenpairs for a large matrix is a daunting task. Archaeologists do not need to worry about how to do it, as many computer math-libraries provide appropriate tools.

As part of this project, we have implemented a tool-box to perform spectral clustering.

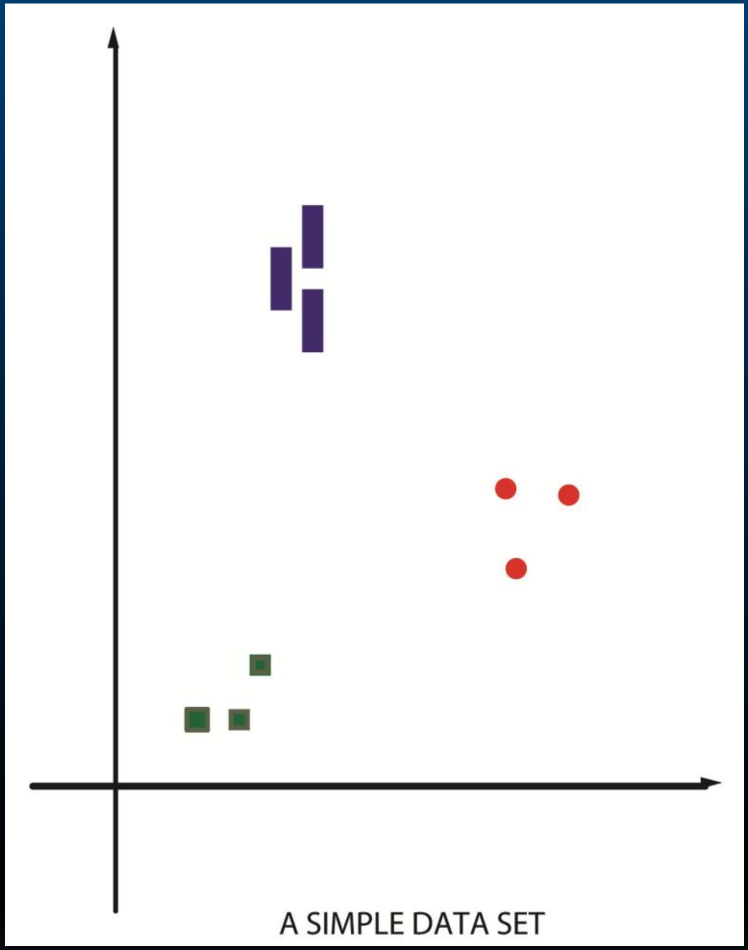
A very basic example

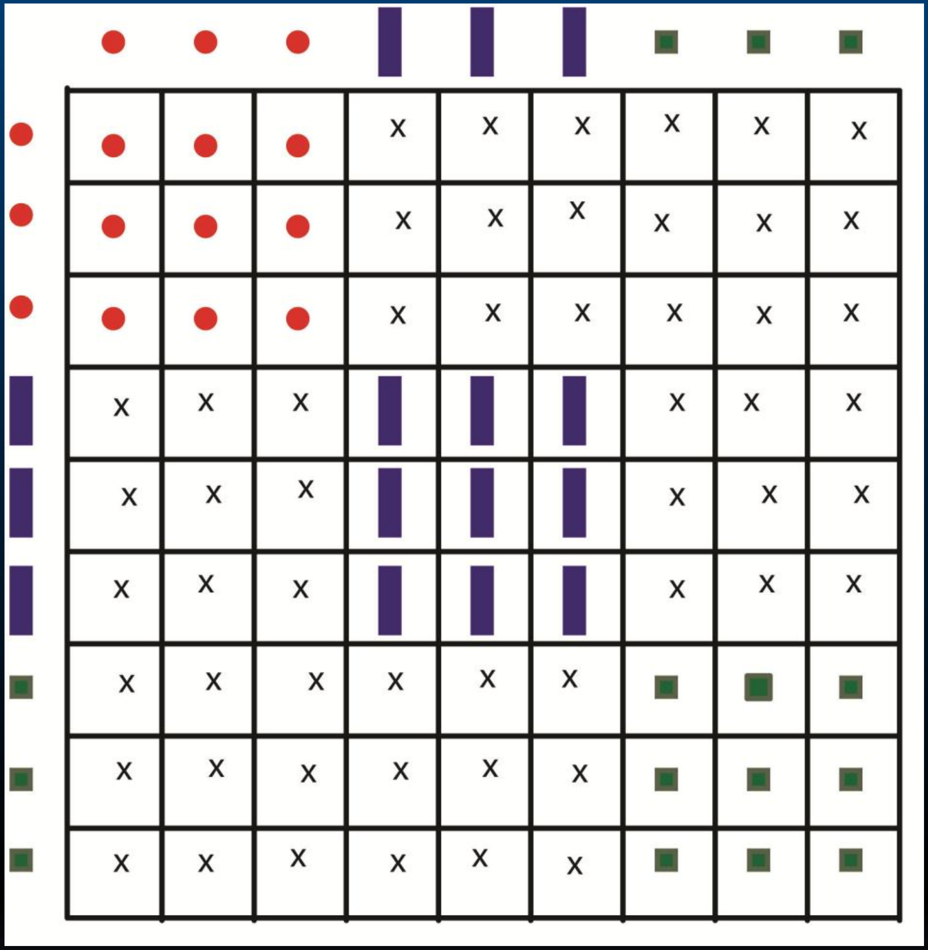
- We try to partition a data set consisting of 9 objects. We first built a 9 x 9 square table, in which both rows and columns enumerate each one of the 9 objects of this example. Then, we calculate the affinity between the objects by applying equation 1 and input the resulting values into the table.

$$S(A,B) = exp\left\{-\frac{1}{2\sigma^2}\left(\frac{1}{n}\sum_{n=1}^{n} I(a_i \neq b_i)\right)^2\right\}$$



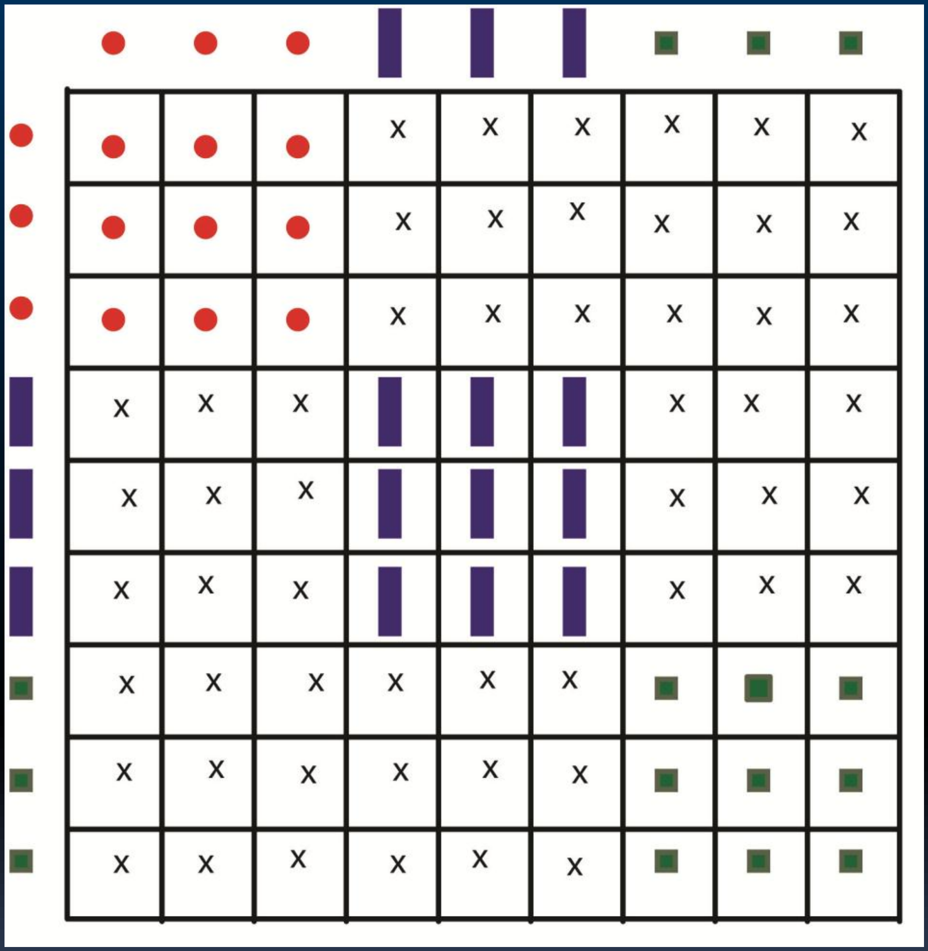
A SIMPLE DATA SET

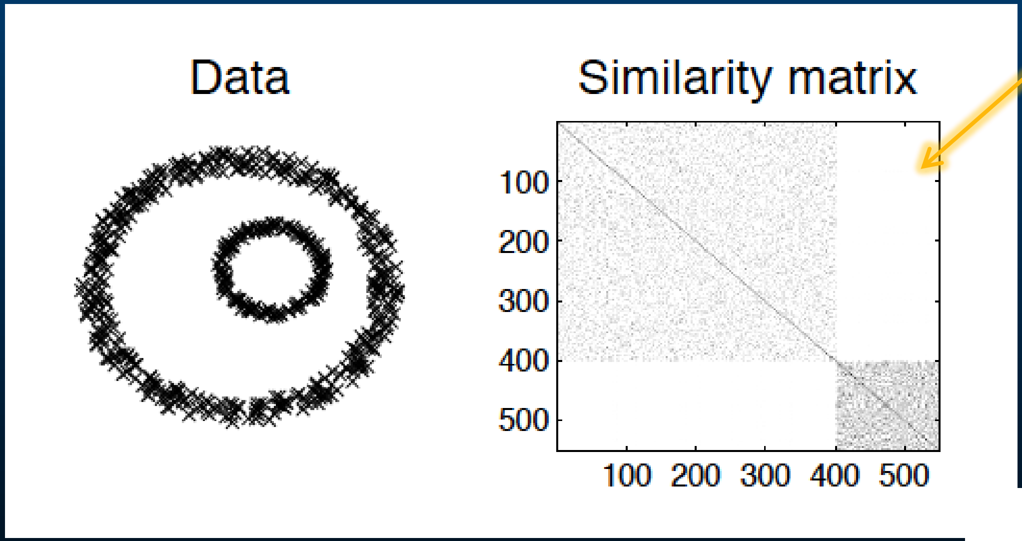




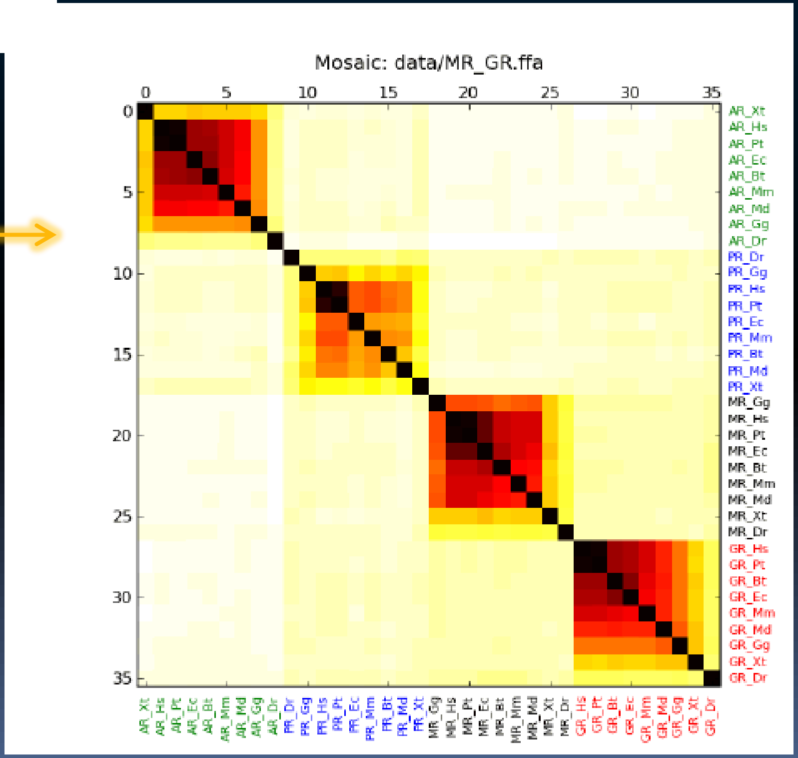
If we codify affinity values with shades of color, then the affinity matrix will have bolder colors in the cells with high affinity and light or no color in those with low affinity.

- The interesting thing to notice is that columns 1, 2, and 3 of the affinity matrix look exactly the same.
- An obvious conclusion is that objects belonging to the same class (cluster) will have similar affinity vectors, which will be quite different of the affinity vectors of other classes.





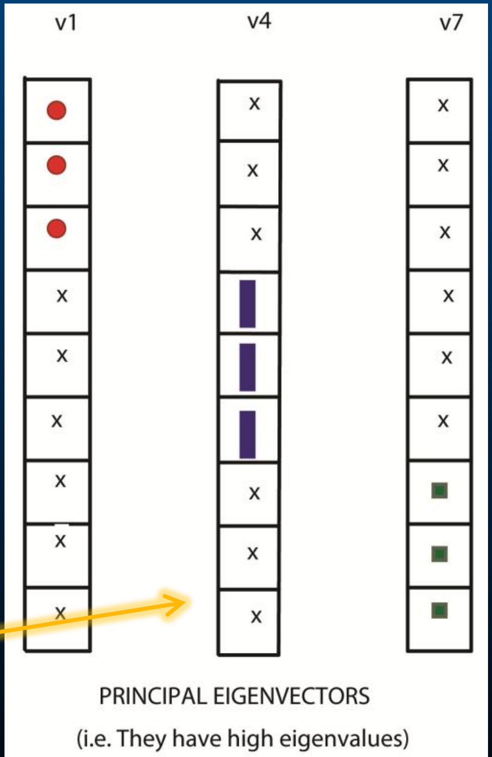
In this picture the affinity matrix shows clearly 2 clusters in the data.



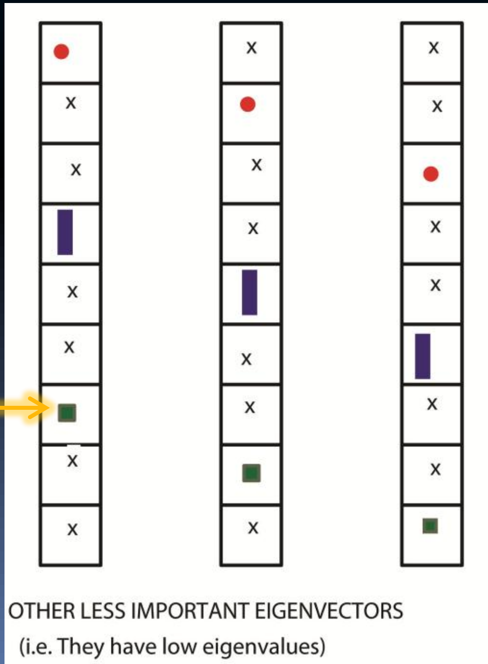
In this picture the affinity matrix shows 4 clusters in the data.

Therefore, by finding the characteristic vectors in the matrix we would be able to identify the clusters in a collection.

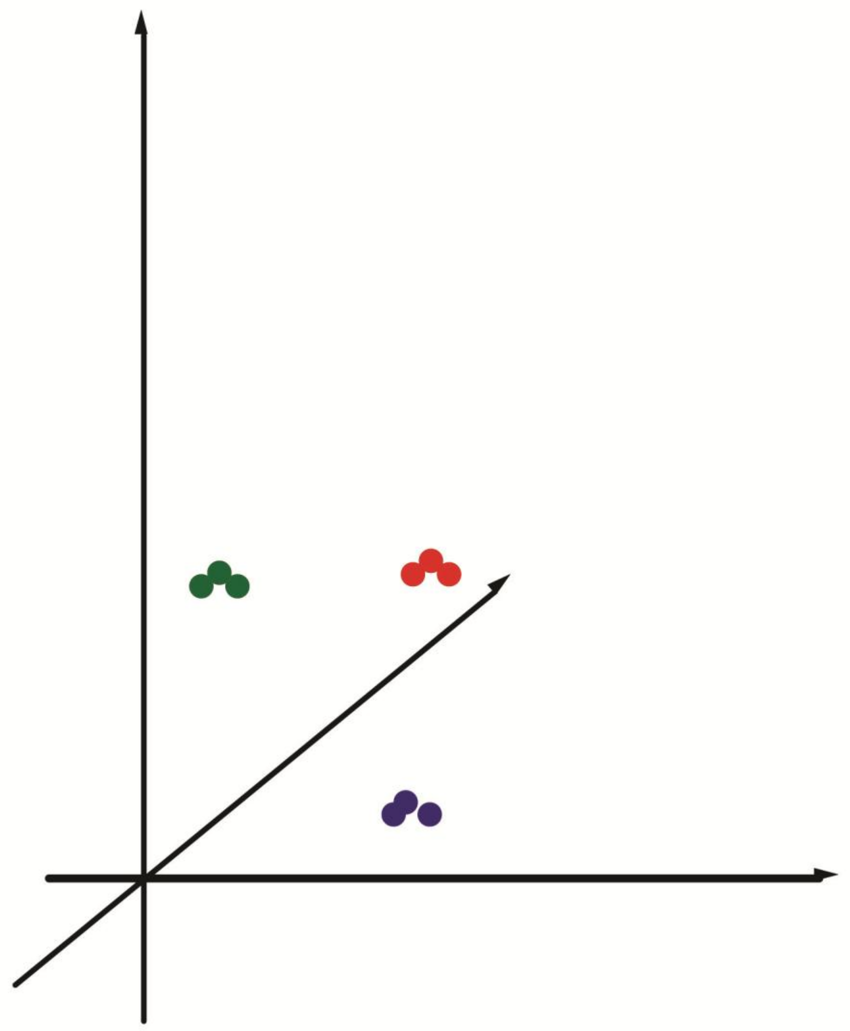
In our example, the three dominant vectors would be the ones illustrated here. These are known in mathematical terms as the eigenvectors of the matrix.



There are other vectors, but these tell us nothing about the clustering structure of the data.



Finally, if we use those vectors as axes of a new coordinate system and map the original data-points into such space, then we would visually appreciate the 3 clusters in a more easy way.



SPECTRAL CLUSTERING

Of course, not all data sets are as clearly structured as the example. In most cases, it is necessary to perform the eigenstructure analysis using complex Linear Algebra algorithms.

Details of those techniques can be found in the extensive literature on the subject, especially in Alpert, Kahng, and Yao (1999), Shi and Malik (2000), Ng, Jordan and Weiss (2001), Melia and Shi (2001), Zelnik-Manor and Prona (2004), Bach and Jordan (2006, 2008), Azranand and Ghahramani (2006b), Yan et al. (2009).

Fortunately, archaeologists do not need to implement these methods, as we have produced a computer program that performs Spectral Clustering automatically.

Results from component linkage

COMPONENT LINKAGE FAILURES

- Dataset of 162 stone masks
- 40 clusters (too many!)
- 10 well defined clusters (too few)
- 6 clusters not very well defined
- Sparse clusters: 13 clusters have only 2 elements.
- 20 masks are not clustered (i.e. 20 clusters have only 1 element).

Component Linkage: Cluster 2

Component Linkage : Cluster 4

Component Linkage: Cluster 10



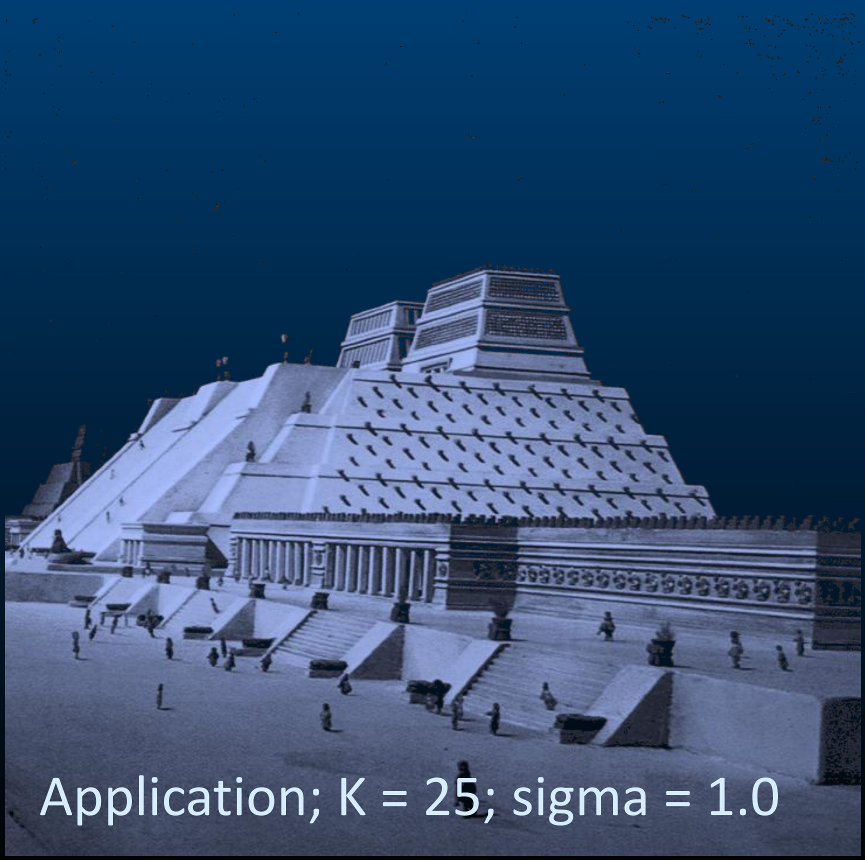



Component Linkage: Cluster 37 (acceptable)

Again, Cluster 37 (as it should be)

Component Linkage: Cluster 39

If spectral clustering is efficient and robust, we should find better-defined clusters



Results from spectral clustering

Cluster 3






Cluster 11



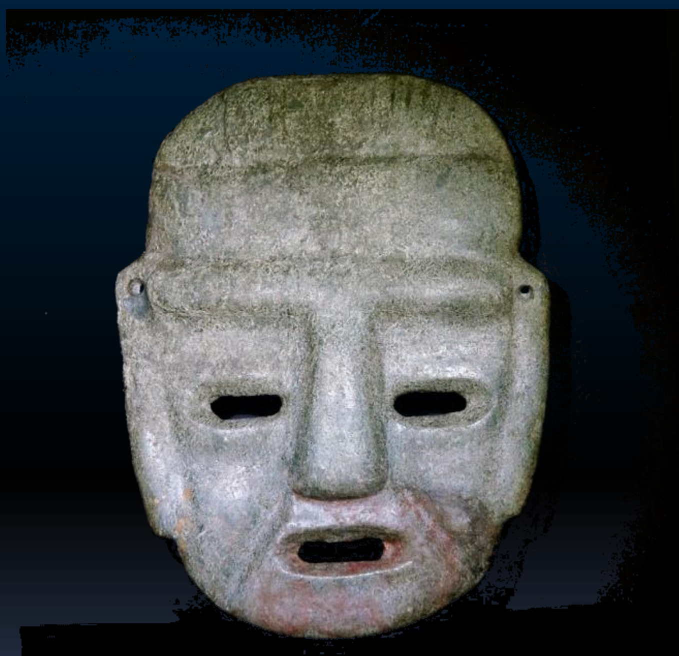




Cluster 23

Cluster 14

Cluster 22



Cluster 18

Cluster 20

Cluster 4

Cluster 25

FUTURE WORK

- Spectral clustering relies on two parameters

1. k (desired number of clusters)
2. σ (similarity threshold)

- Our next goal is to apply another algorithm to “learn” the value of k directly from the dataset.

- Software demonstration

Conclusions

- Applying Spectral Clustering to the Mezcala collection have produced encouraging results. We were able to partition the mask collection into 23 well-defined groups, which is a better result than the 40 clusters obtained with Numerical taxonomy.
- We illustrated the 23 groups of Mezcala masks. The reader would notice the great performance of the Spectral Clustering algorithm, especially by comparing clusters 3, 11, 12, 14, 16, 18, 25, and some others. Such groups are defined by highly similar masks. Cluster 12, for example, contains masks with triangular faces made in highly polished stone. In contrast, cluster 3 contains square masks, most of which with perforated eyes and less polished material that the ones in group 12.

- Furthermore, each one of the groups identified with Spectral Clustering is clearly different from the rest, which allows us trusting the partition. Only 2 masks (labeled here as “clusters” 7 and 17) were left un-clustered, which represents a better result than the one obtained by numerical taxonomy in which 20 masks were isolated.
- Therefore, we believe that Spectral Clustering may have a future role in archaeology, especially as a first step in analyzing shape features of complex collections.

THANKS FOR YOUR ATTENTION