IPAM October 2, 2007

Multi Scale organizations in diffusion geometry of digital documents. Data driven "ontology/language".

The Harmonic Analysis of "folders" and observables on networks

Ronald Coifman

Program of Applied Mathematics ,Yale University

Joint work with:

J.Bremer, P.Jones,S. Lafon, M. Maggioni, B. Nadler, F. Warner, Y. Keller, A. Singer, Y Shkolnisky, Y. Kevrekides, S.W. Zucker.

We elaborate on the idea that "*The Network*" encapsulates knowledge.

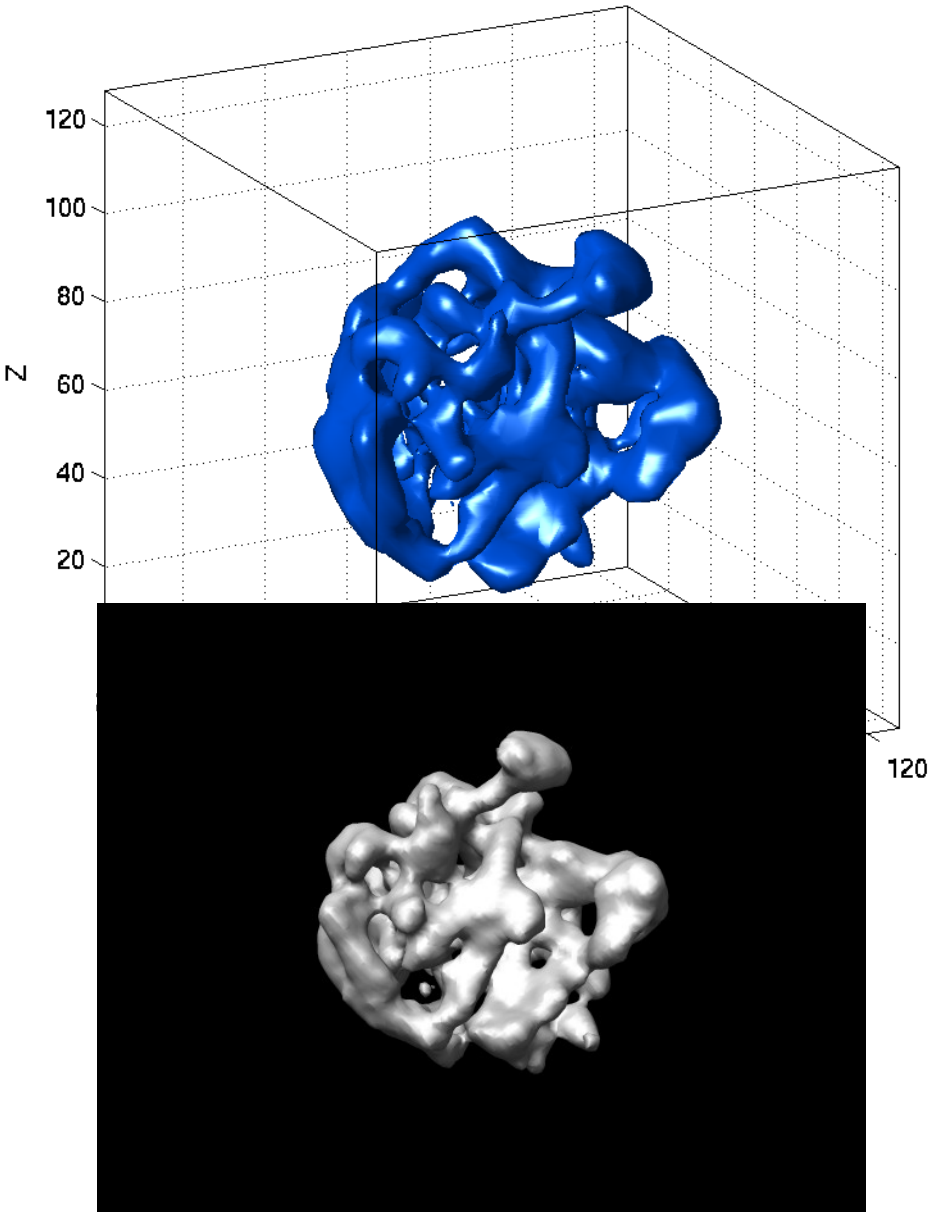
- Mathematical analysis of recent algorithmic insights in Data organization and Search engines enable the formalization of the concept that knowledge can be encapsulated by the geometry of the "web/network" of connections and affinity relations between complex data strings.
- The "social networks" of inferences between digital documents, text, proteins ,sensor, characterize their function .
- A diagnostic estimate could be viewed as a function on the "geometric network" of such configurations, and could be expanded in basis functions on the network

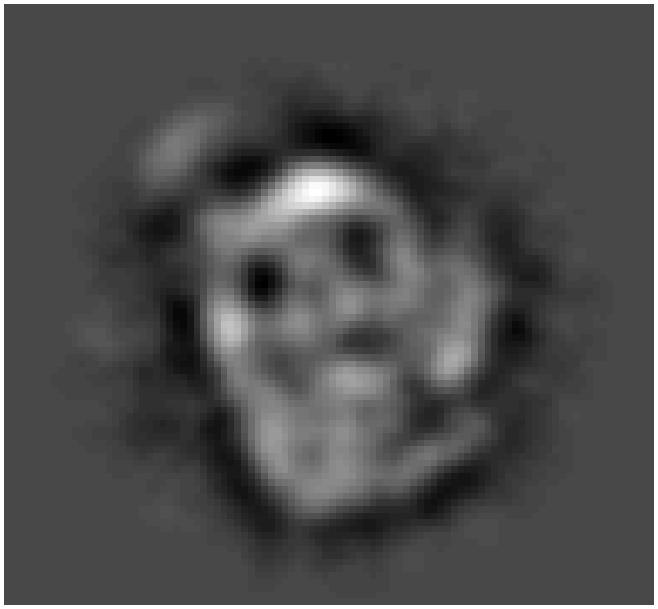
This simple point is illustrated below

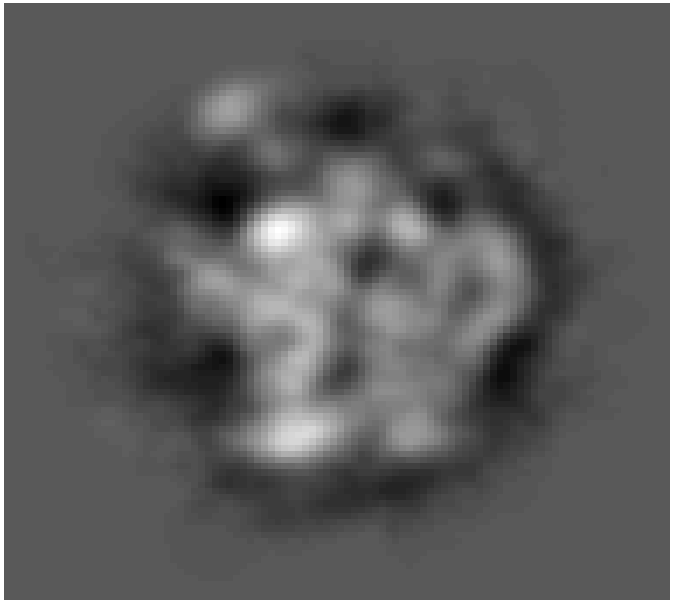
Three Dimensional Puzzle

Each puzzle piece is linked to its neighbors (in feature space) the network of links forms a sphere. A parameterization of the sphere can be obtained from the eigenvectors of the inference matrix relating affinity links between pieces (diffusion operator).

Cryo-microscopy is an example of the spherical puzzle, the orientation of a molecular image is unknown, and is being determined through the graph of similarities







A simple empirical diffusion/inference matrix A can be constructed as follows

Let X_i ***represent normalized data ,we "soft truncate" the covariance matrix as***

$$A_0 = [X_i \bullet X_j]_\varepsilon = \exp\{-(1 - X_i \bullet X_j)/\varepsilon\}$$

$$\|X_i\| = 1$$

A is a renormalized Markov version of this matrix

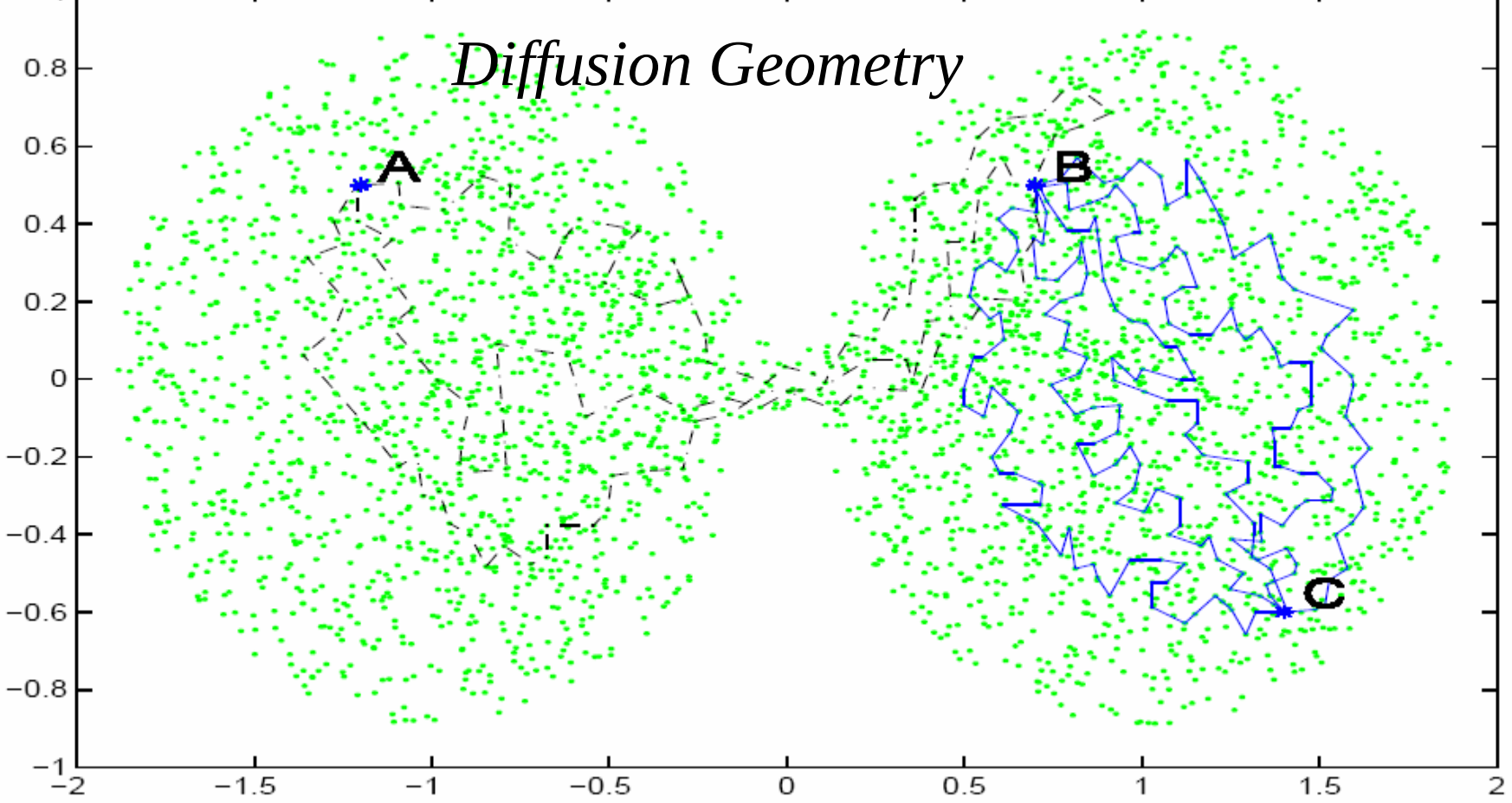
The eigenvectors of this matrix provide a local non linear principal component analysis of the data . Whose entries are the diffusion coordinates These are also the eigenfunctions of the discrete Graph Laplace Operator.

$$A^t = \sum \lambda_l^{2t} \phi_l(X_i)\phi_l(X_j) = a_t(X_i, X_j)$$

$$X_i^{(t)} \to (\lambda_1{}^t \phi_1(X_i), \lambda_2{}^t \phi_2(X_i), \lambda_3{}^t \phi_3(X_i), ..)$$

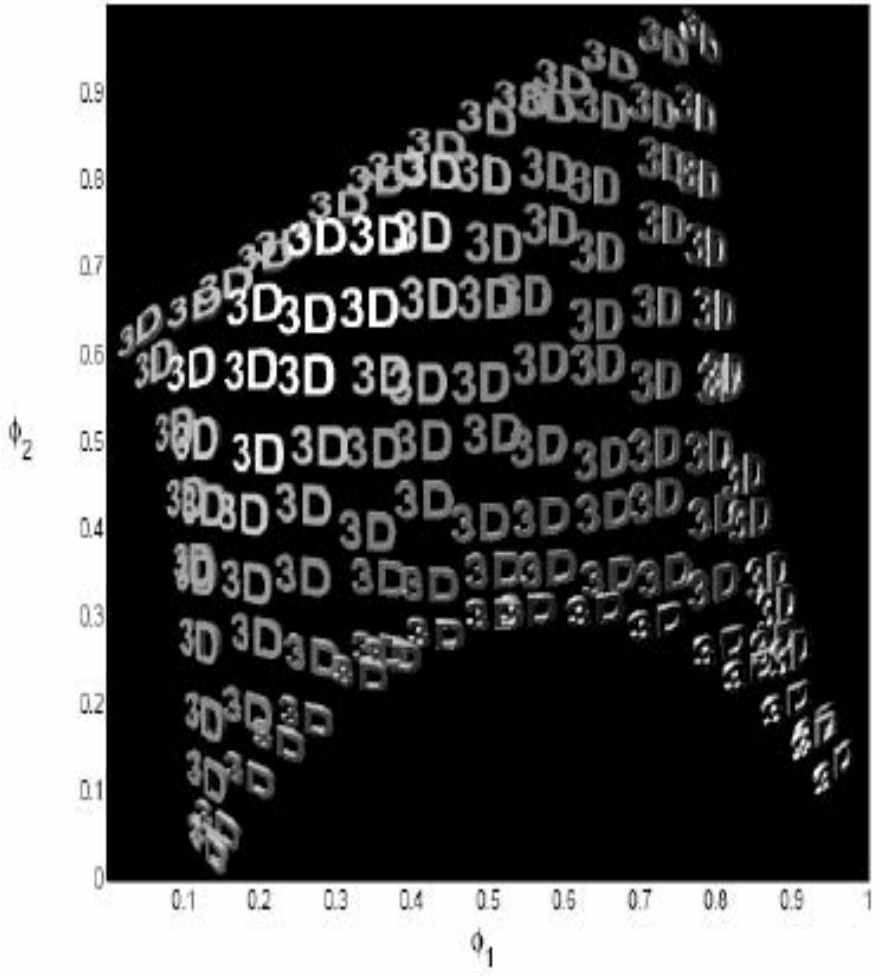
$$d_t^2(X_i, X_j) = a_t(X_i, X_i) + a_t(X_j, X_j) - 2a_t(X_i, X_j) = \left\| X_i^{(t)} - X_j^{(t)} \right\|^2$$

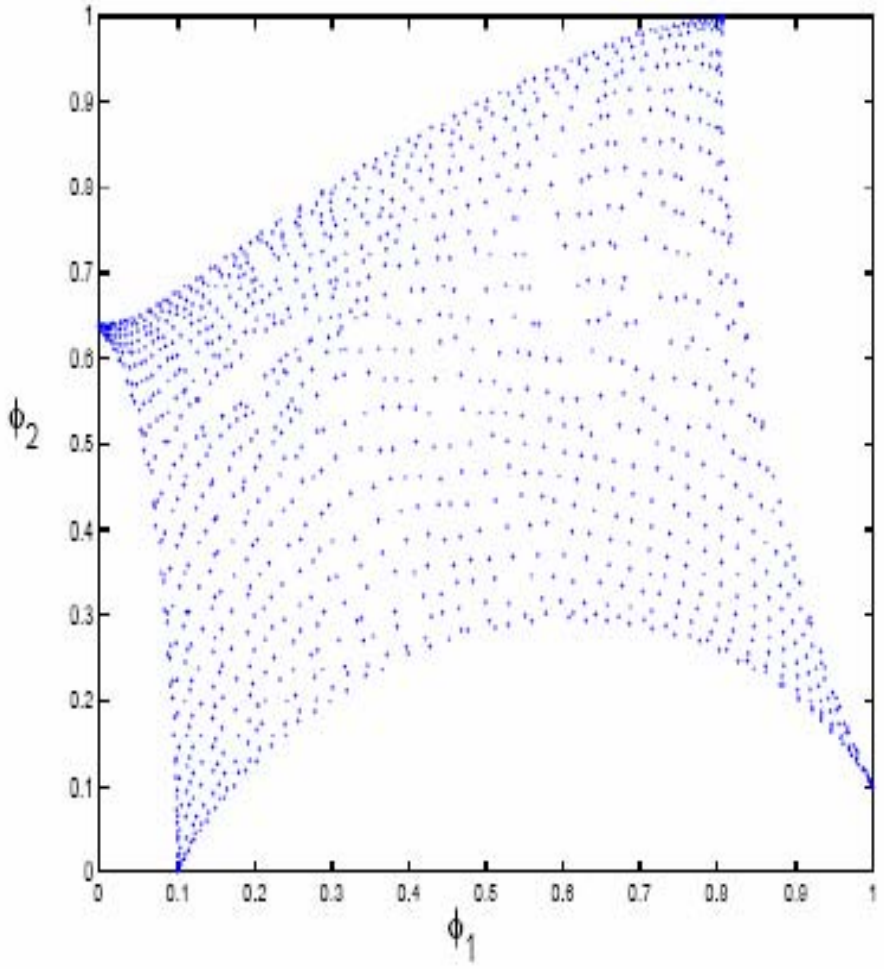
This map is a diffusion embedding into Euclidean space (at time t) .



Diffusions between A and B have to go through the bottleneck ,while C is easily reachable from B. The Markov matrix defining a diffusion could be given by a kernel , or by inference between neighboring nodes.

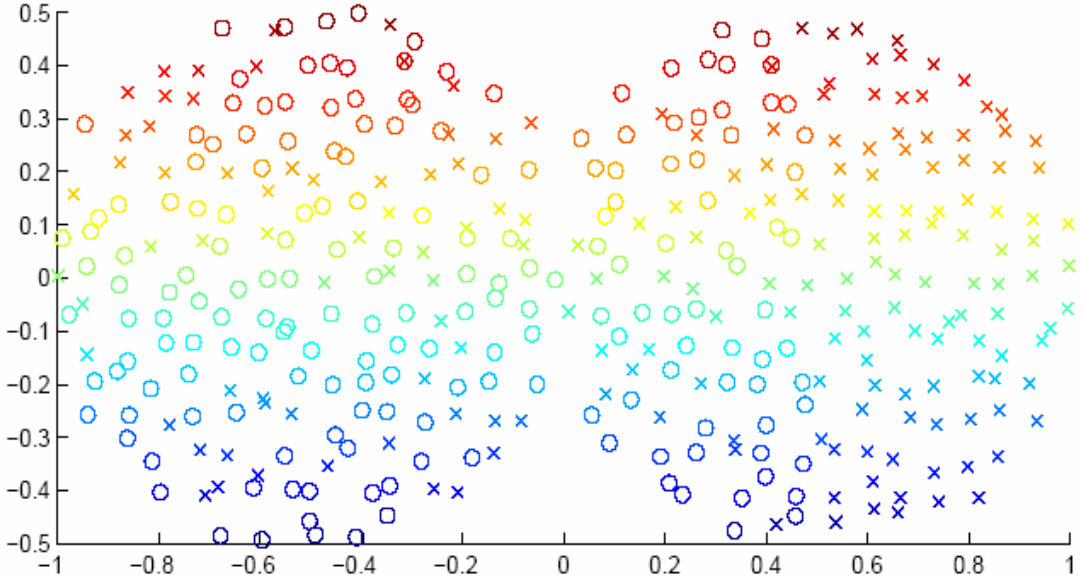
The diffusion distance accounts for preponderance of inference links . The shortest path between A and C is roughly the same as between B and C . The diffusion distance however is larger since diffusion occurs through a bottleneck.

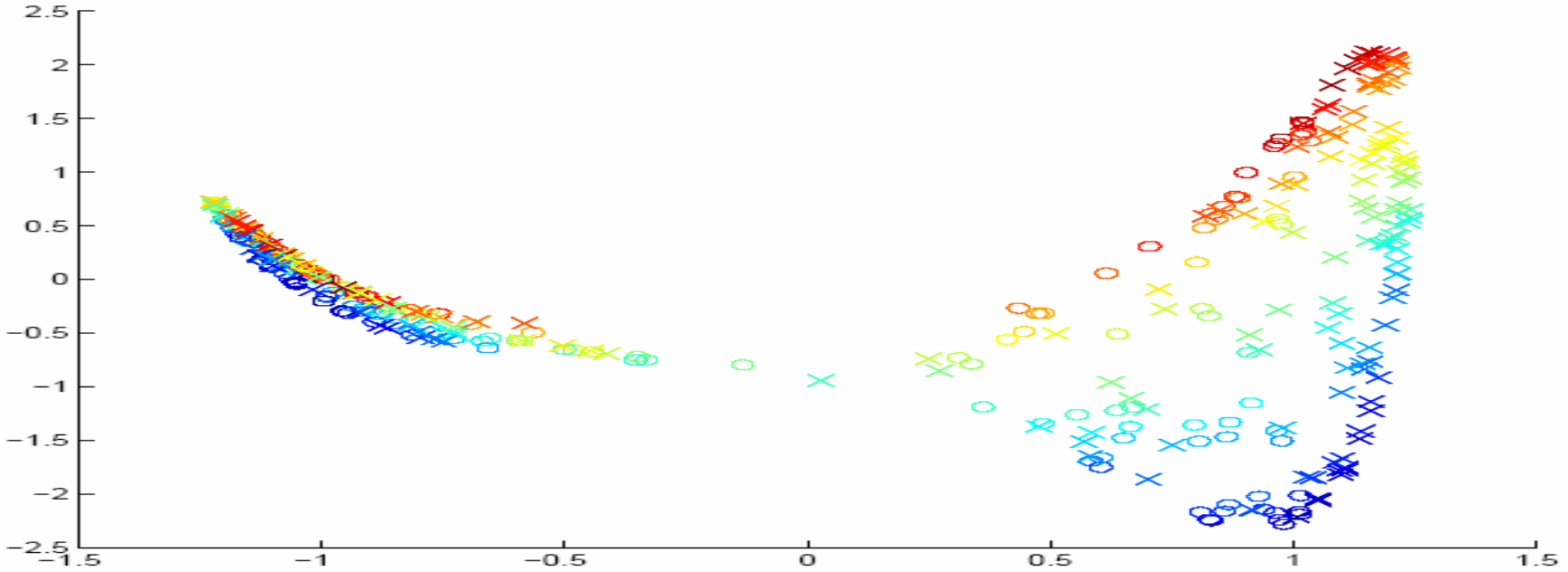


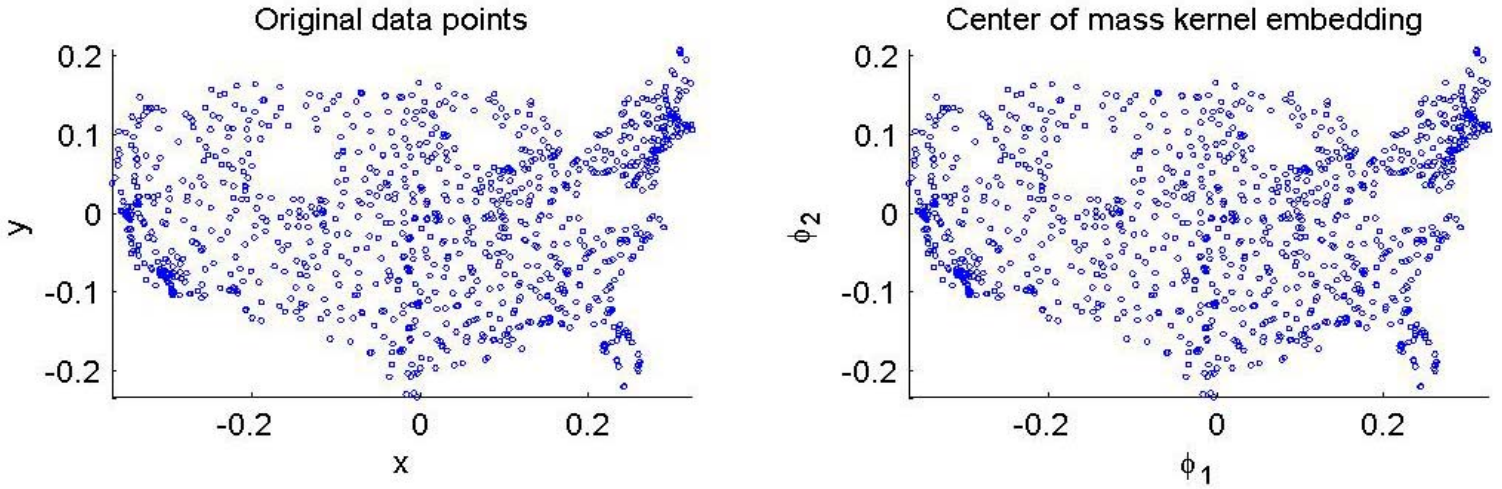


The First two eigenfunctions organize the small images which were provided in random order, in fact assembling the 3D puzzle.

The long term diffusion of heterogeneous material is remapped below . The left side has a higher proportion of heat conducting material ,thereby reducing the diffusion distance among points , the bottle neck increases that distance







Local information such as the distance between nearby cities can be encapsulated in the local center of mass matrix , whose eigenvectors include the x and y coordinate functions (see A. Singer)

$$\sum_{j=1}^{k} W_{i,i_j} r_{i_j} = r_i$$

$$\sum_{j=1}^{k} W_{i,i_j} = 1$$

Paradigm blend

Classical Harmonic Analysis through Calderon Zygmund theory has been concerned with organization and approximation of functions and operators through a blend of Combinatorics, Geometry and Fourier modes (which are eigenfunctions of Laplaceans).

It turns out that in the data context a similar methodology blending geometry and harmonics is as useful.

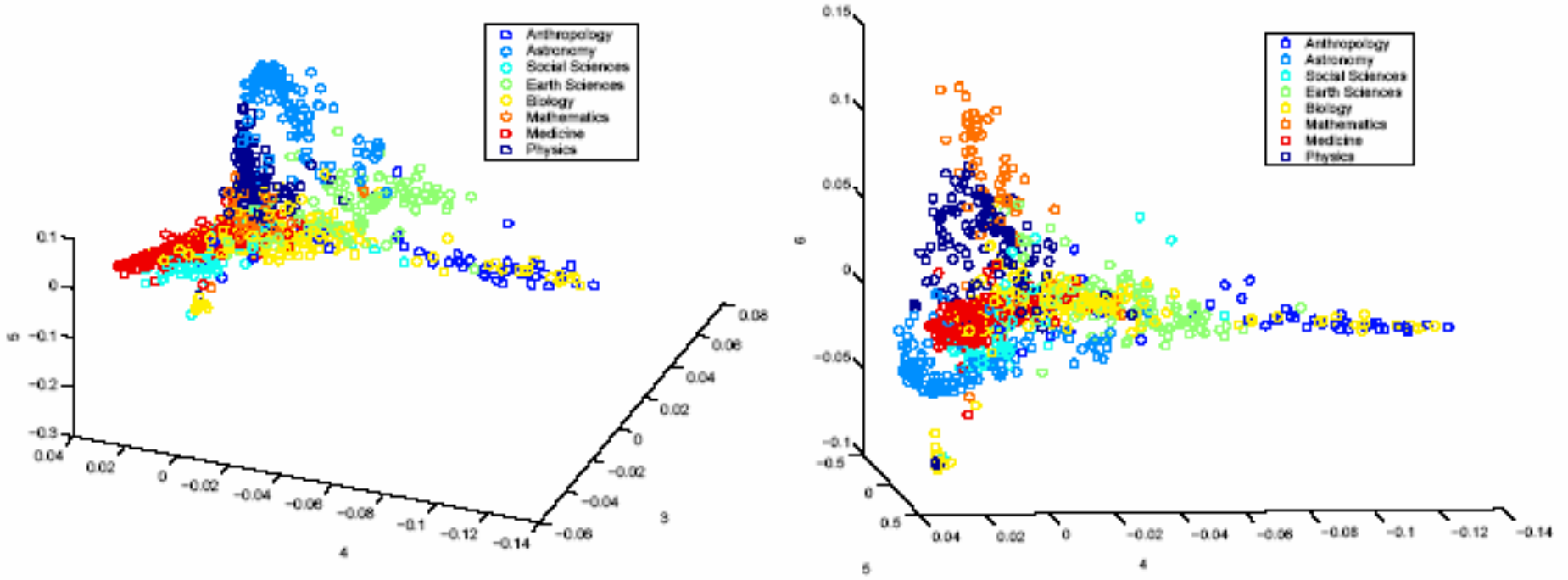
Consider the “Globe puzzle” the eigenfunctions of the appropriate affinity are the solution to the puzzle . On the other hand we would solve the puzzle by first piecing together nearest pieces (in features) , and proceed by linking such patches by their own affinity etc at several scales.

This method turns out to be mathematically equivalent to the Eigenfunction paradigm , in the same way that wavelets relate to Fourier series.

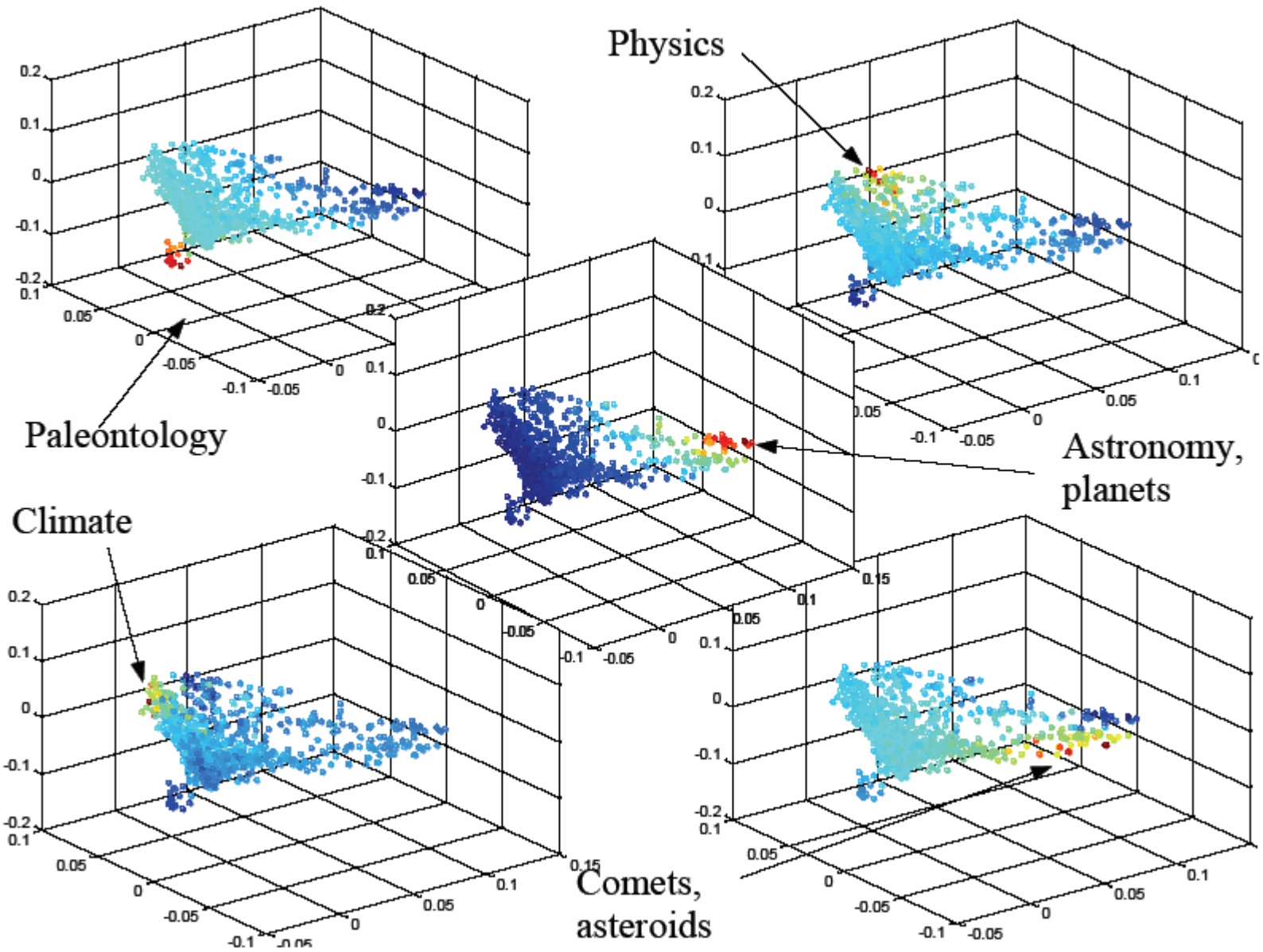
Here organization is achieved through ,eigenfunctions and wavelet constructions

Application to text document classification

1000 Science News articles, from 8 different categories. We compute about 10000 coordinates, i-th coordinate of document d represents frequency in document d of the i-th word in a fixed dictionary. The diffusion map gives the embedding below. Clustering in the range of diffusion map results in good unsupervised performance for document classification.

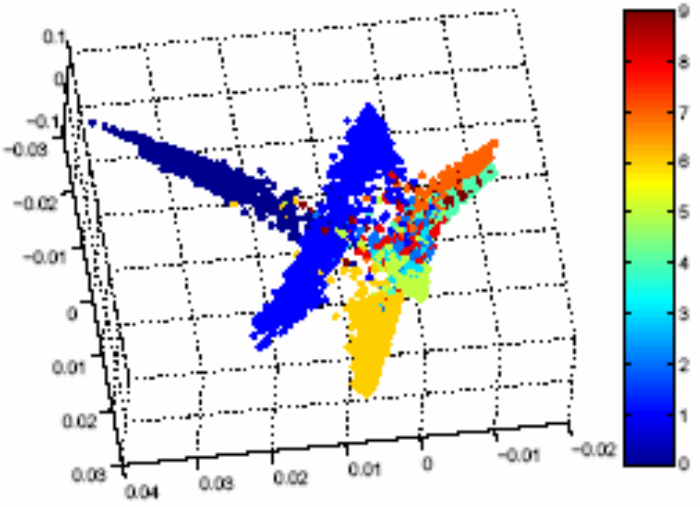


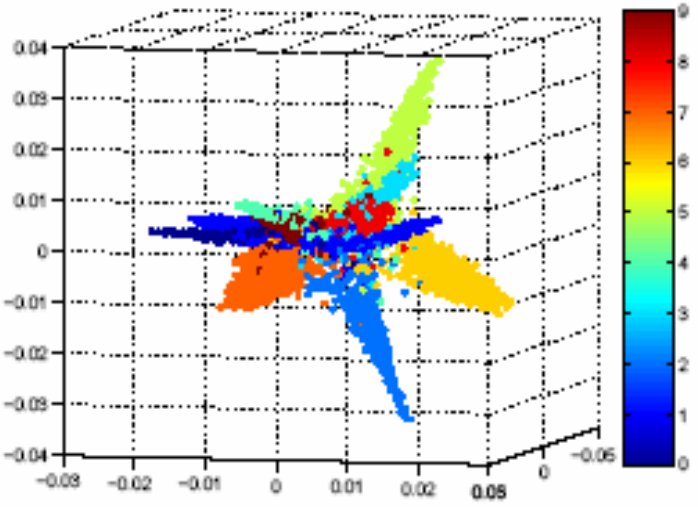
Embedding $\Xi_6^{(0)}(x) = (\xi_1(x), \ldots, \xi_6(x))$: on the left coordinates $3, 4, 5$, and on the right coordinates $4, 5, 6$.


Physics
Paleontology
Astronomy, planets
Climate
Comets, asteroids

Handwritten Digits

Data base of about $60,000$
28×28 gray-scale pictures of handwritten digits, collected by USPS. Goal: automatic recognition.
It is a point cloud in 28^2 dimensions. We can think of being given this cloud, and some points are labeled by the digit they correspond to, and we would like to predict the digit corresponding to each point.





Set of $10,000$ picture (28 by 28 pixels) of 10 handwritten digits. Color represents the label (digit) of each point.

Multiscale organization of Graphs.

We now describe a simple booking strategy to organize folders on a data graph. We follow the “puzzle strategy”

We organize a graph into a hierarchy of graphs consisting of disjoint subsets at different time scales of diffusion.

Let

$a_t(x,y)$ be the diffusion at time t on the graph,

i, e $a_t(x,y)$ is the kernel of the power t of the diffusion operator

$$A^1(f)(x) = \int a_1(x,y) f(y) dy$$

$$d_t^2(x,y) = a_t(x,x) + a_t(y,y) - 2a_t(x,y)$$

is the distance at scale t between x and y,

This distance can be extended to define a distance between subsets by

$$d_{2t}^{\ 2}(E,F) = \int \left| \int a_t(x,y)[\chi_E(y) - \chi_F(y)]dy \right|^2 dx =$$

$$= a_{2t}(E,E) + a_{2t}(F,F) - 2a_{2t}(E,F)$$

where

$$a_t(E,F) = \iint a_t(x,y)\chi_F(x)\chi_E(y)dxdy$$

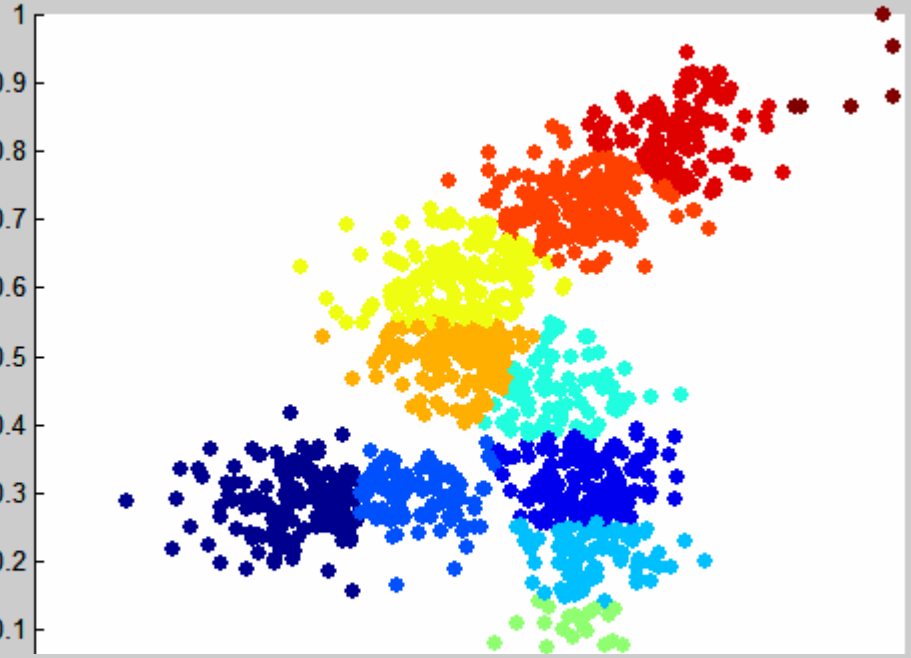
here

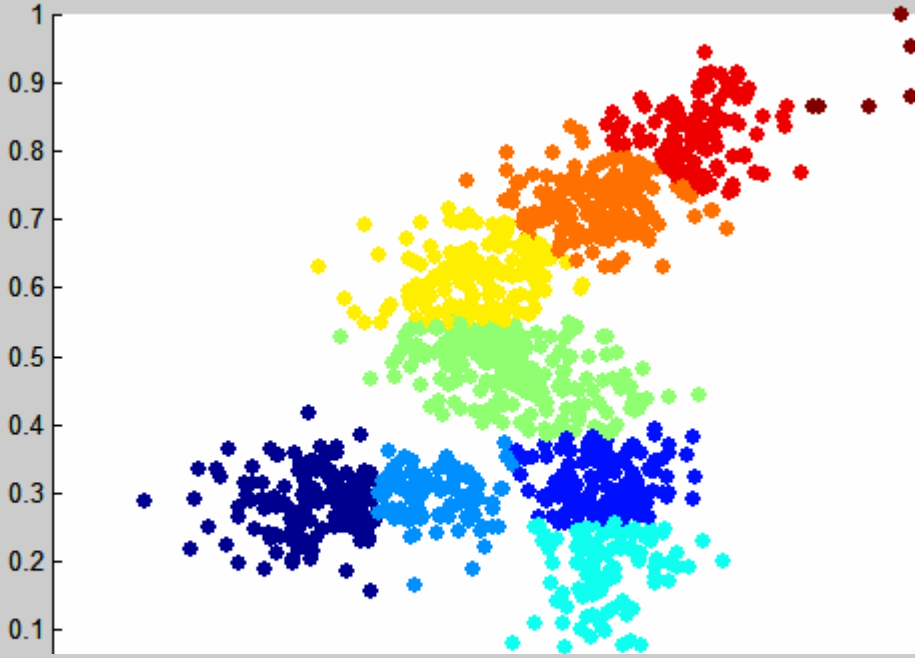
$$\chi_F(x) = \frac{1}{\sqrt{|F|}}$$ when x is in F and 0 otherwise.

A very simple way to build a hierarchical multiscale structure is as follows.

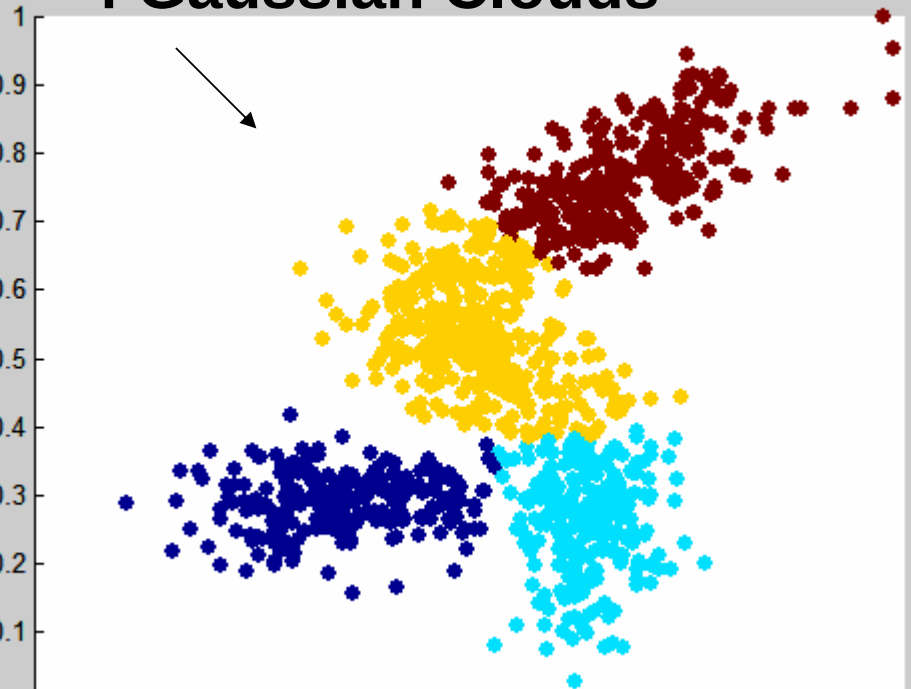
Start with a disjoint partition of the graph into clusters of diameter between 1 and 2 relative in the diffuion distance with t=2.
Consider the new graph formed by letting the elements of the partition be the vertices .
Using the distance between sets and affinity between sets described above we repeat with t=4, until we end with one folder, and a tree of graphs ,each a coarse version of the preceding with its own temporally rescaled geometry (folder structure)

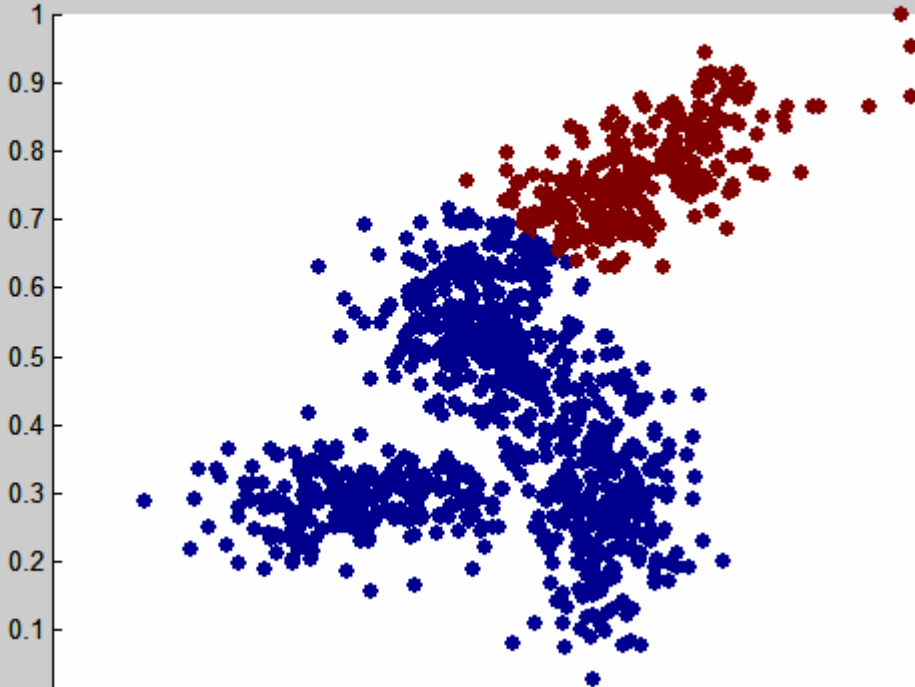
In the next image we see this organization as it applies to a random collection of 4 Gaussian clouds .

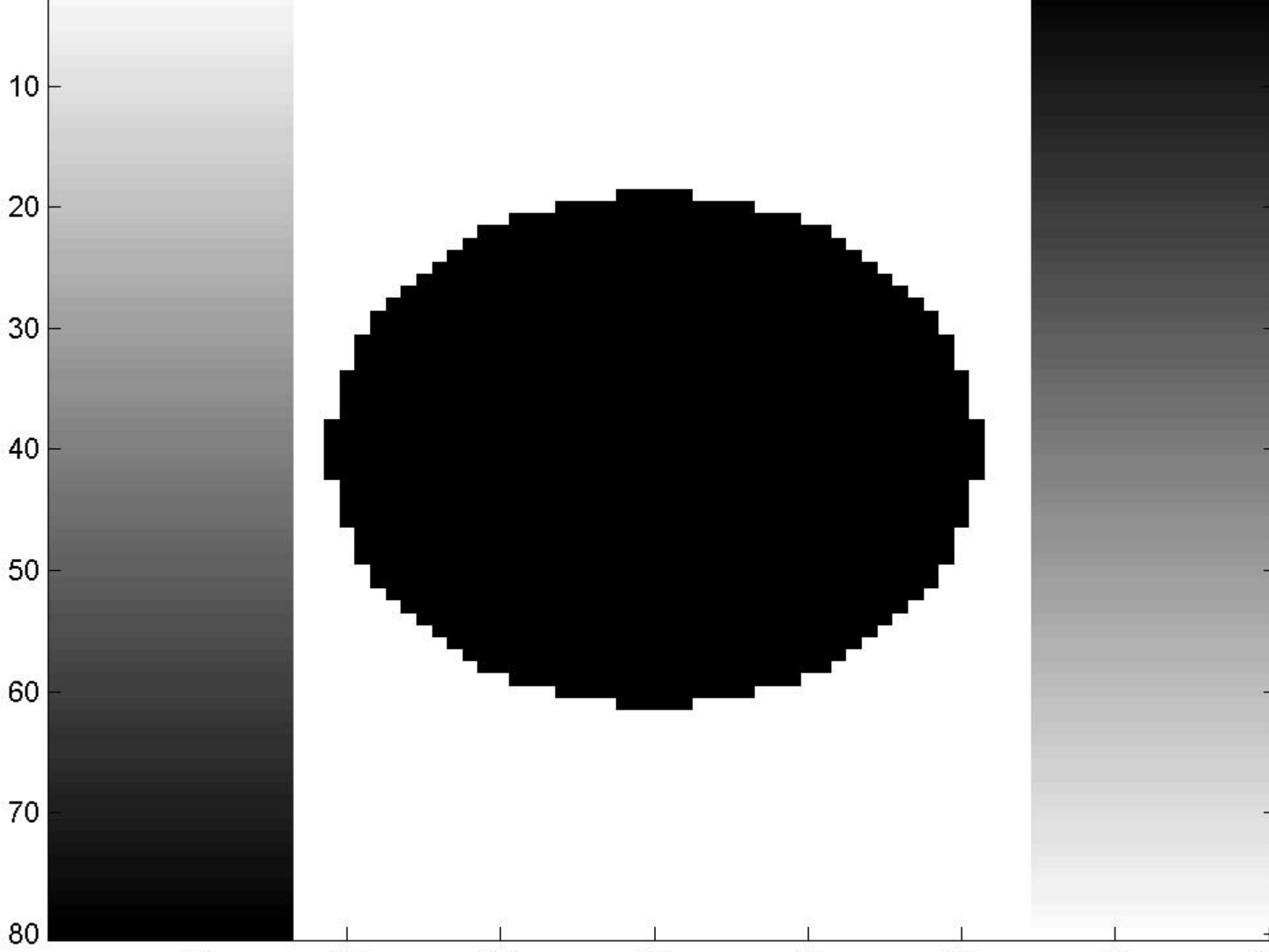




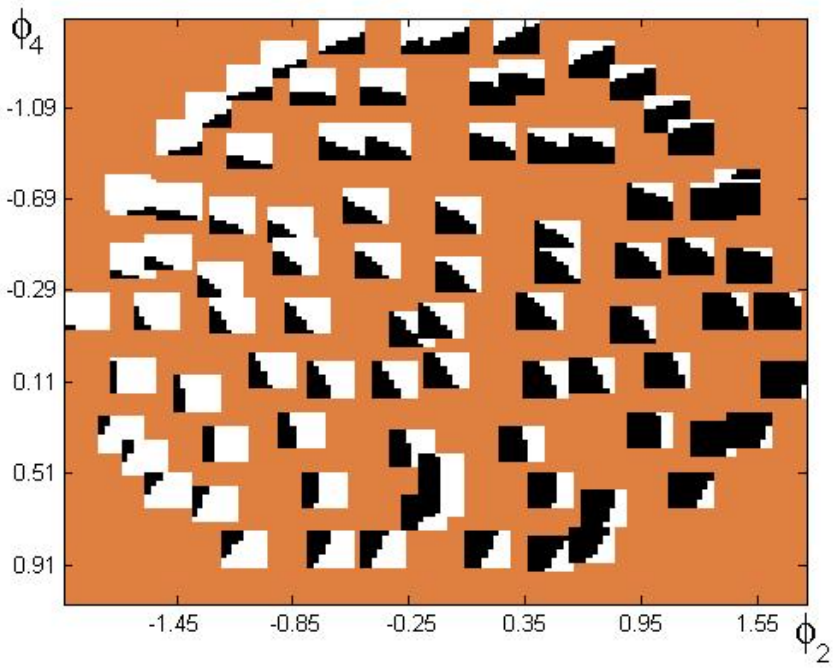
4 Gaussian Clouds



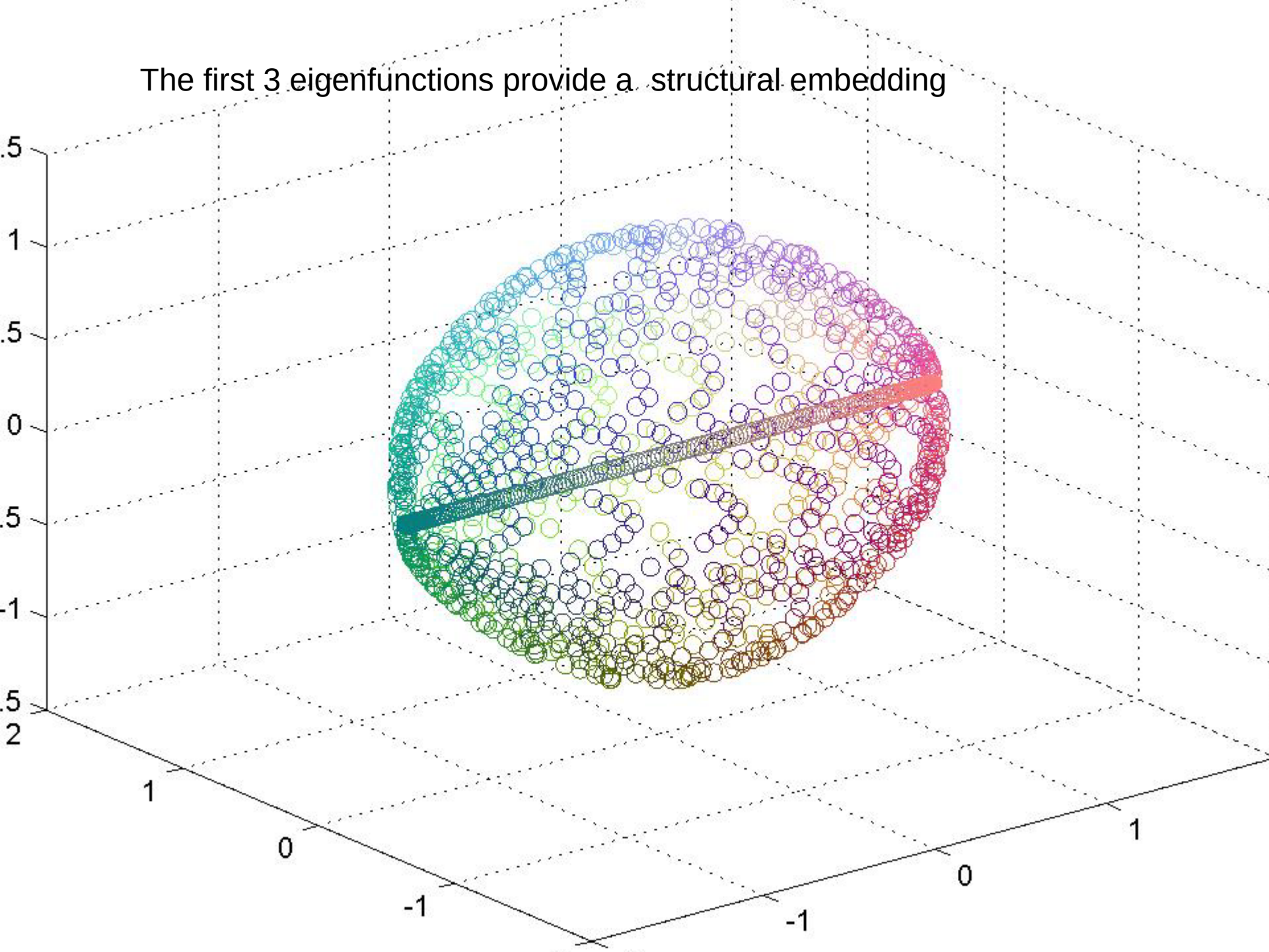


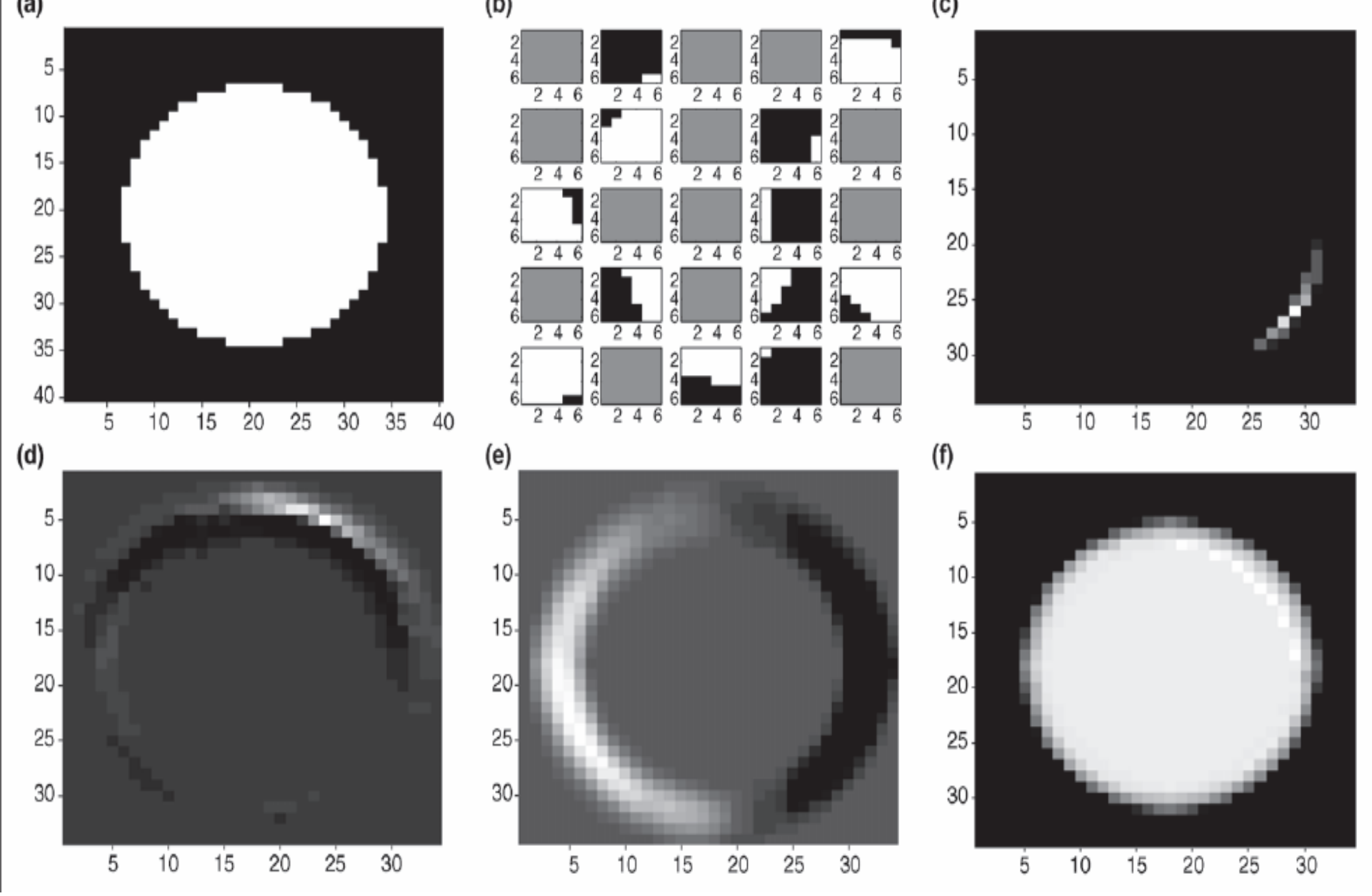

10
20
30
40
50
60
70
80

We now organize the set of subimages of 8x8 squares extracted from the preceding image and organized naturally by their average and orientation of the edge (the first two eigenfunction coordinates) .



The first 3 eigenfunctions provide a structural embedding

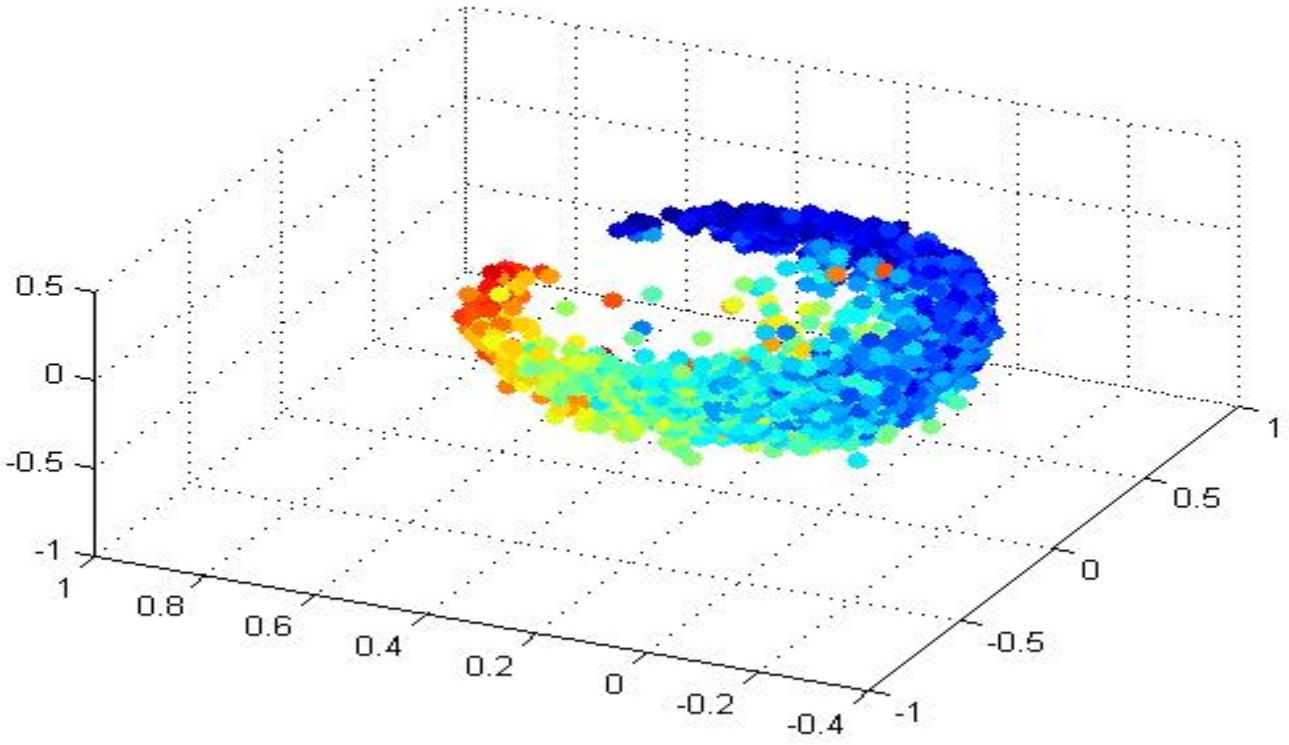




The clusters of nearby points in the multiscale hierarchy ,corresponds ot features in the original image.

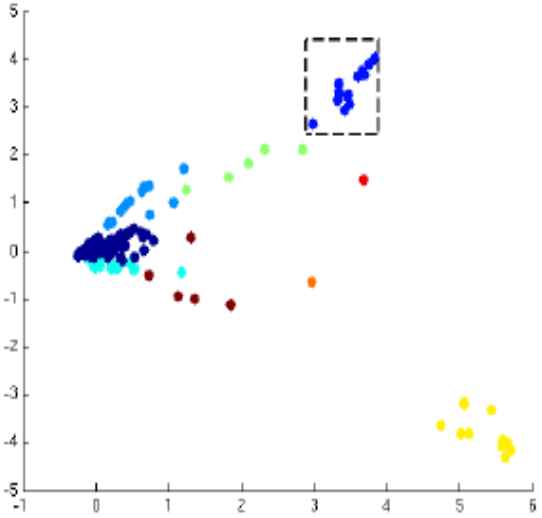
The demographic geometry of responders to the MMPI questionnaire . (500 questions 300 responders)

Each point represents a responder ,nearby people had similar response profile. The color represents a depression score , red is high.

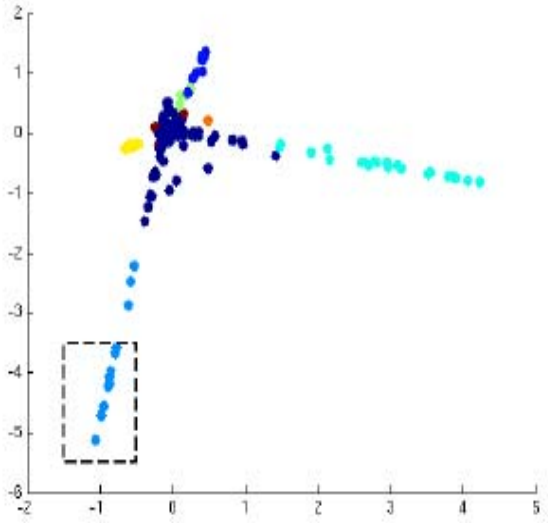


The organization of the questions into a graph leads to topical folders.

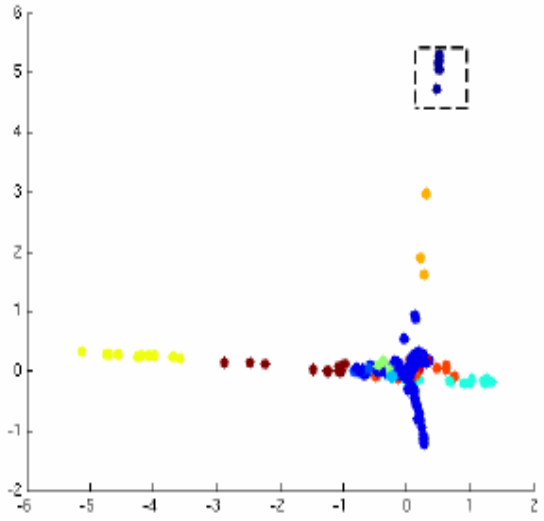
Question Embedding Projections



- I would like to be a nurse.
- If I were a reporter, I would very much like to report news of the theater.
- I used to like to play hopscotch and jumprope.
- I have often wished that I were a girl (or if I am a girl, I enjoy being a girl).



- I have no fear of water.
- I have a great fear of snakes.
- I am not afraid of mice.
- I do not worry about catching diseases.



- Sometime I get so excited that I find it hard to sleep.
- I have periods in which I feel extremely cheerful without any explanation.
- I have periods were I feel so full of pep that sleep is not necessary for days at a time.

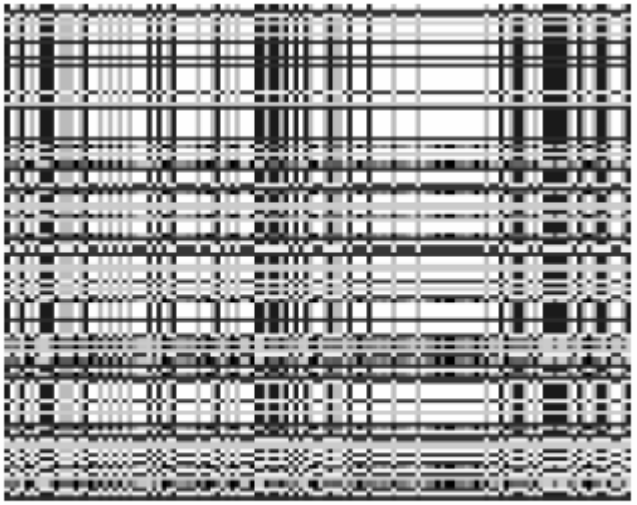
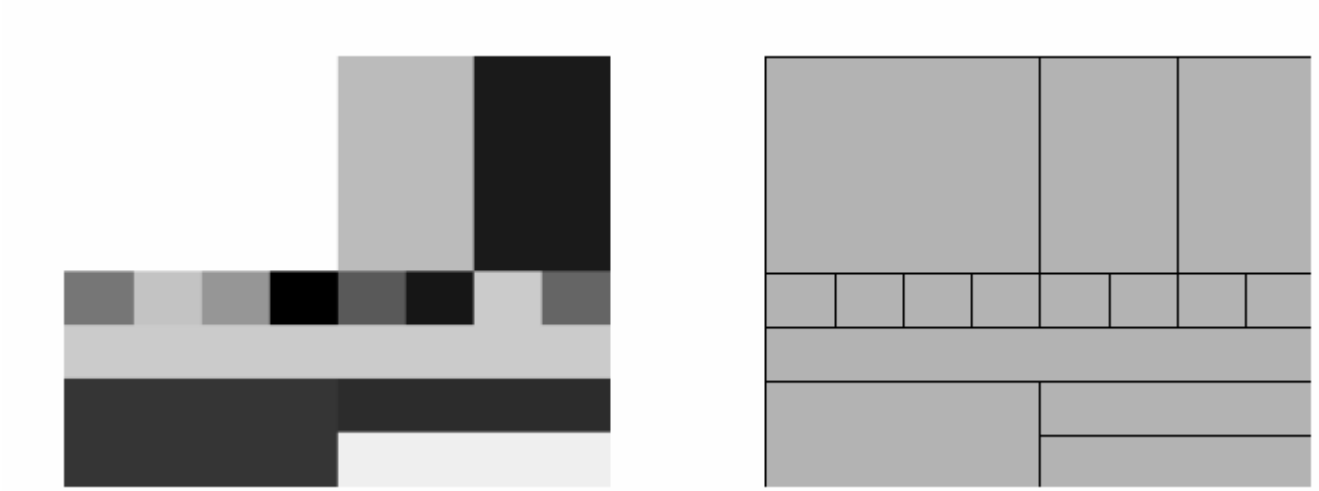
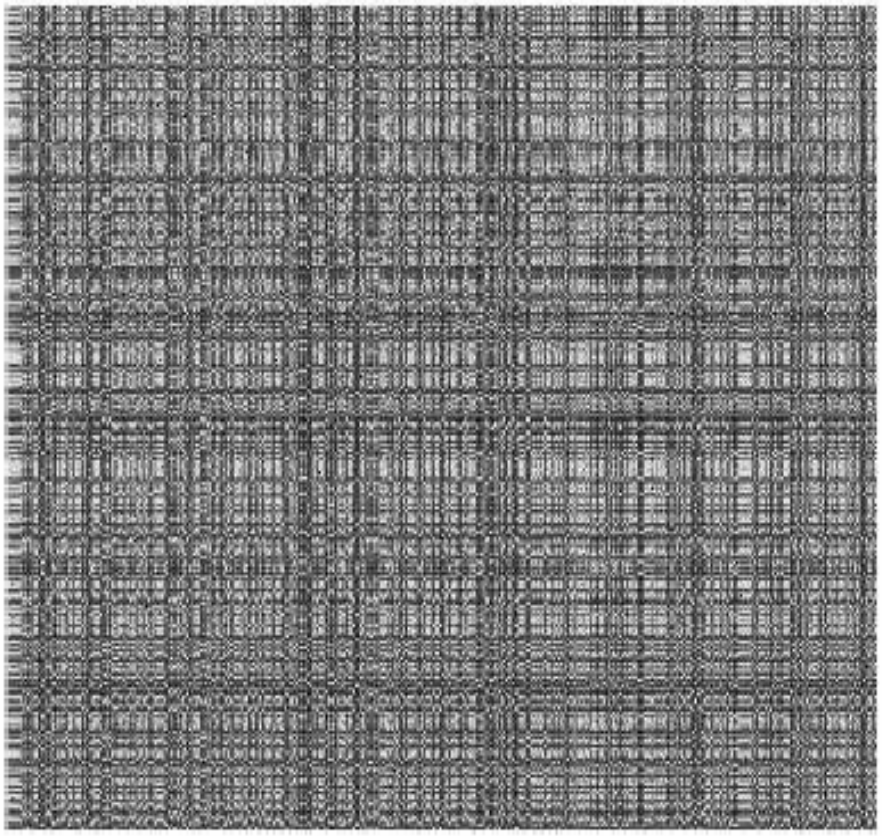


Figure 4.2: A permutation of the matrix A.



Unraveling a simple response matrix (bottom left) which has been permuted.

This is achieved by building row and column geometries of folders and reorganizing to minimize the complexity.



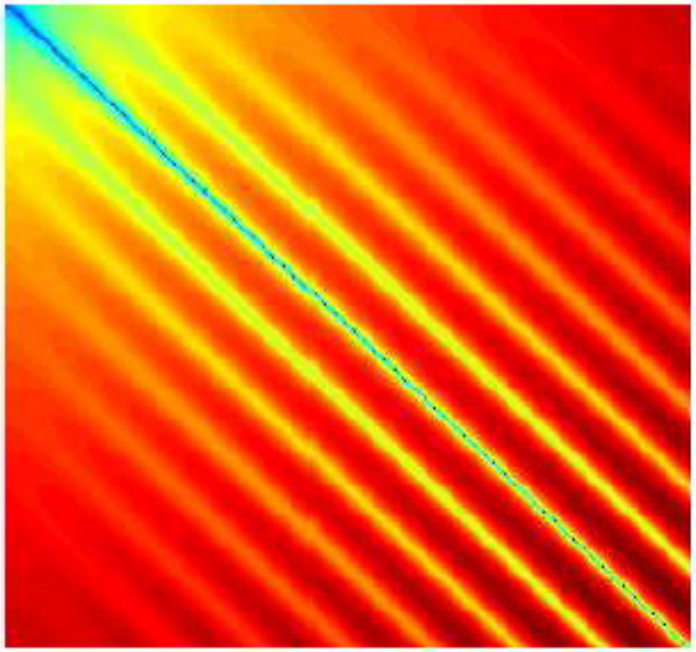
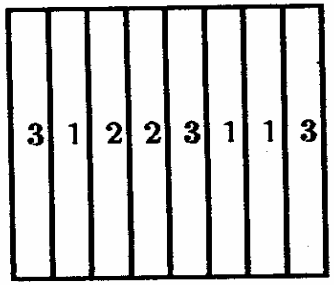
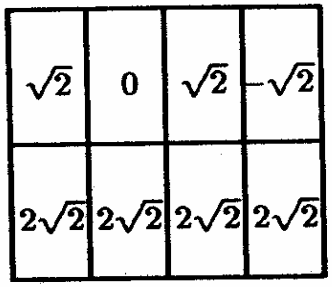
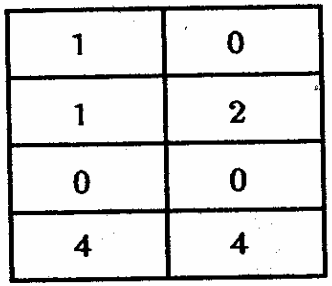
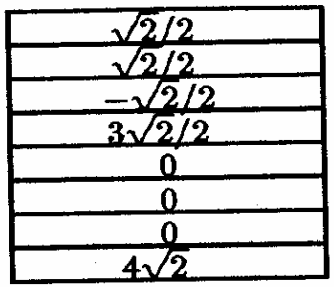


Figure 5.7: The kernel $||x - y||^{1/4}$ on the spiral, before and after permutation.

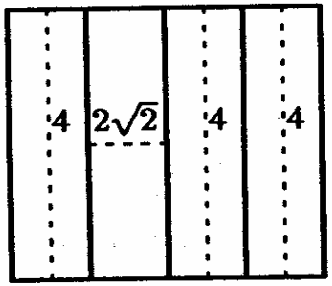

3 1 2 2 3 1 1 3

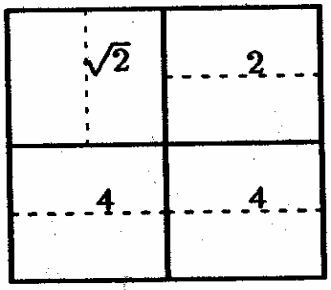

√2 0 √2 −√2
2√2 2√2 2√2 2√2

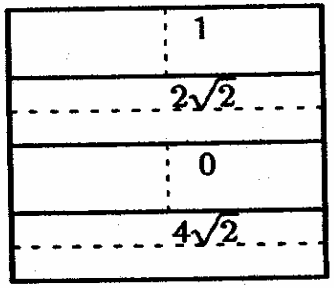

1 0
1 2
0 0
4 4


√2/2
√2/2
−√2/2
3√2/2
0
0
0
4√2

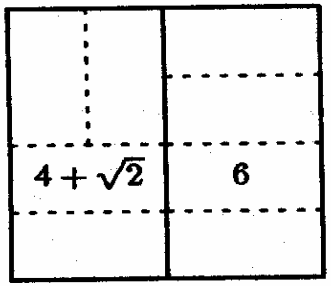
$l = 1$

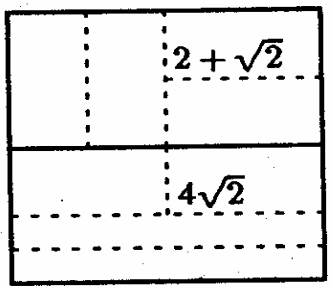

4 2√2 4 4


√2 2
4 4

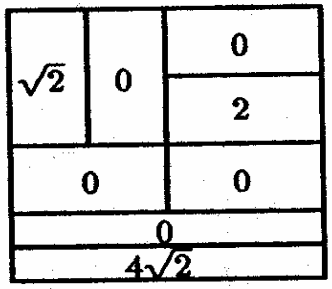

1
2√2
0
4√2

$l = 2$


4 + √2 6


2 + √2
4√2

$l = 3$


√2 0 0
2
0 0
0
4√2

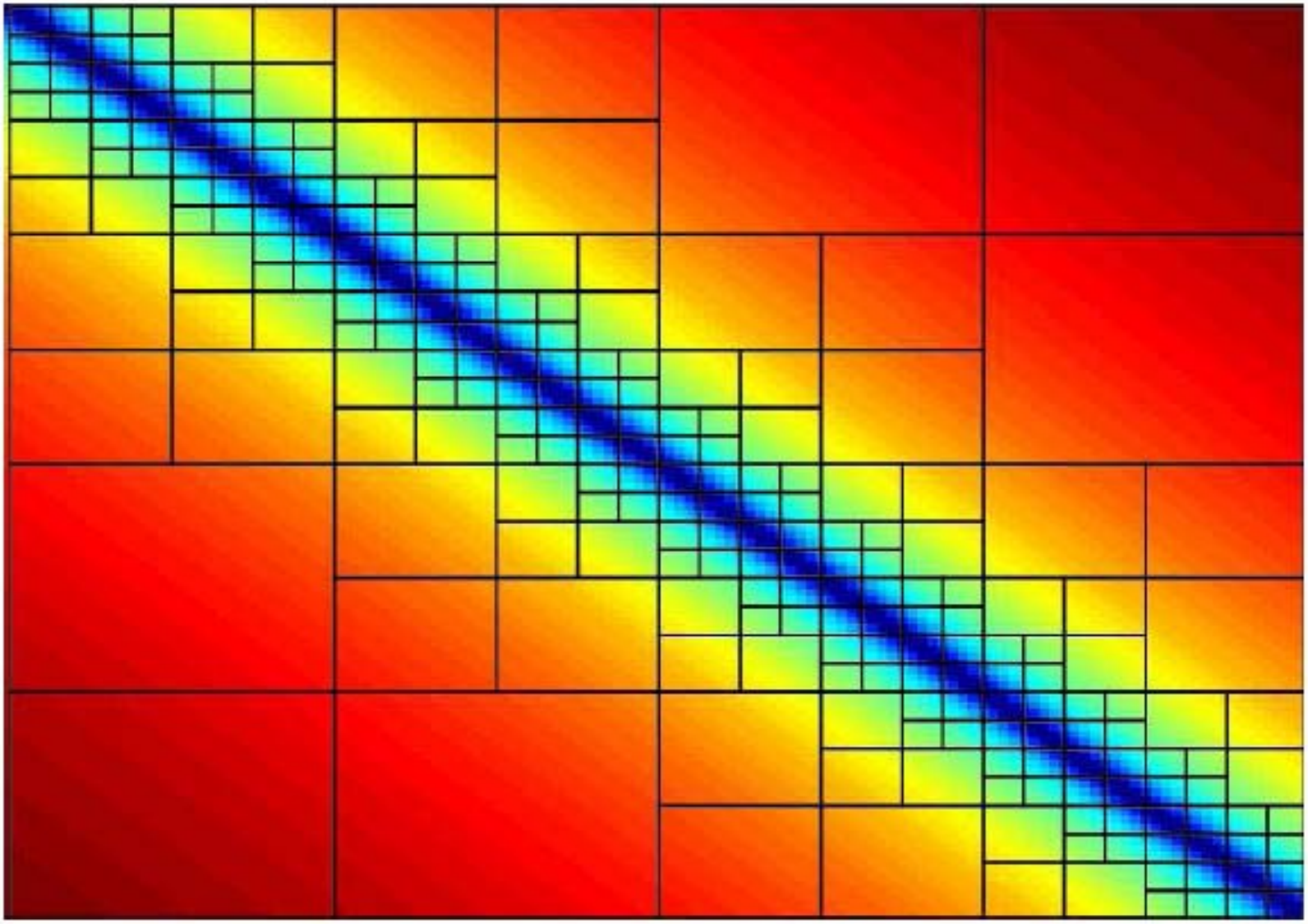
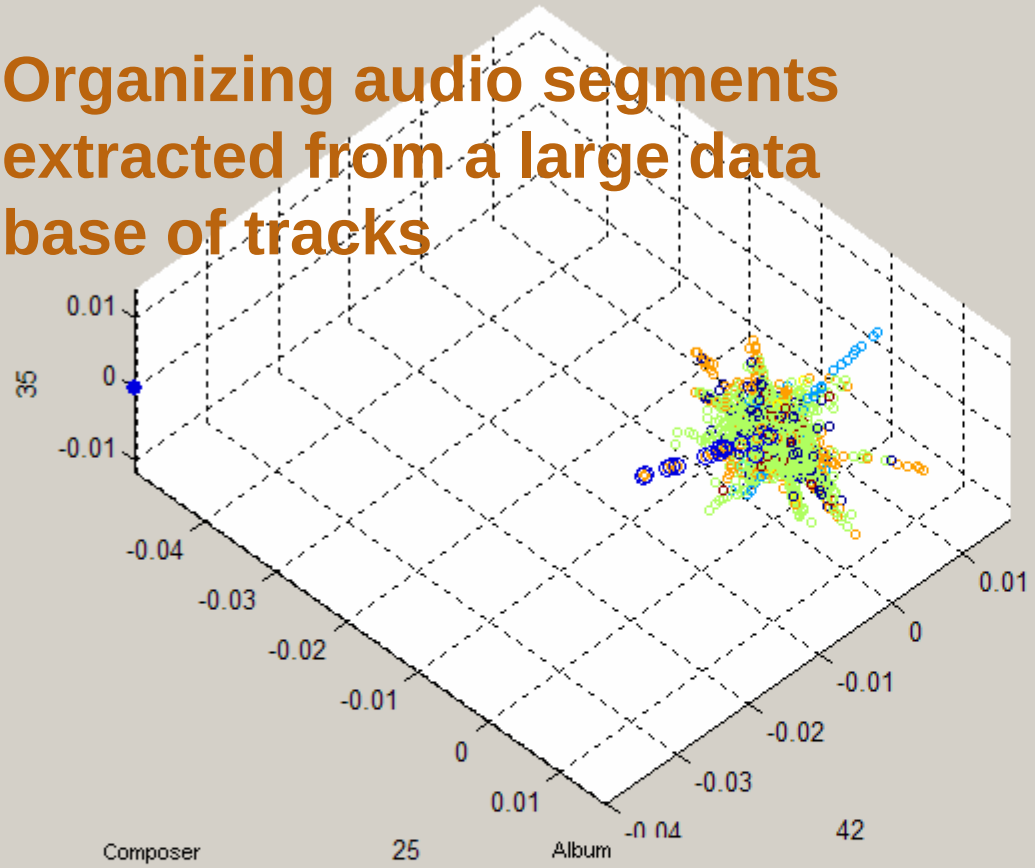


Figure 5.1: The structure of the integral kernel $\frac{1}{|x-y|}$.

GUI_music_02

File

Organizing audio segments extracted from a large data base of tracks



Display options

Scale by eigenvalues

Axis equal

2D

Rotate

Time for diffusion map

Choose axes

25

42

35

Show spectrum

Search Options

Dimensions for distance

Top 100

Number of neighbors for search

Everything But the Girl	Lullaby Of Clubland (CD Maxi Single)	02 - Everything But The Girl - Lullaby Of Clubland (Markus Sc
Everything But the Girl	Lullaby Of Clubland (CD Maxi Single)	02 - Everything But The Girl - Lullaby Of Clubland (Markus Sc
Everything But the Girl	Lullaby Of Clubland (CD Maxi Single)	02 - Everything But The Girl - Lullaby Of Clubland (Markus Sc
Everything But the Girl	Lullaby Of Clubland (CD Maxi Single)	06 - Everything But The Girl - Lullaby Of Clubland (Markus Sc
Underworld	Beaucoup Fish	01 - Underworld - Cups
Everything But the Girl	Lullaby Of Clubland (CD Maxi Single)	02 - Everything But The Girl - Lullaby Of Clubland (Markus Sc
Everything But the Girl	Lullaby Of Clubland (CD Maxi Single)	06 - Everything But The Girl - Lullaby Of Clubland (Markus Sc
Mary J. Blige	Mary	03 - Mary J. Blige - Deep Inside
Massive Attack	Mezzanine	07 - Massive Attack - Man Next Door
Suba	Tributo	07 - Suba - Samba Do Gringo Paulista
The Spinners	One of a Kind Love Affair (2 of 2)	07 - The Spinners - Wake up Susan
Elvis Costello & the Attractions		
Emily Remler		

> PLAY

[] STOP

Make Track New Center

Save Song List

Save MP3s

Search in the whole database

->

->

No track selected

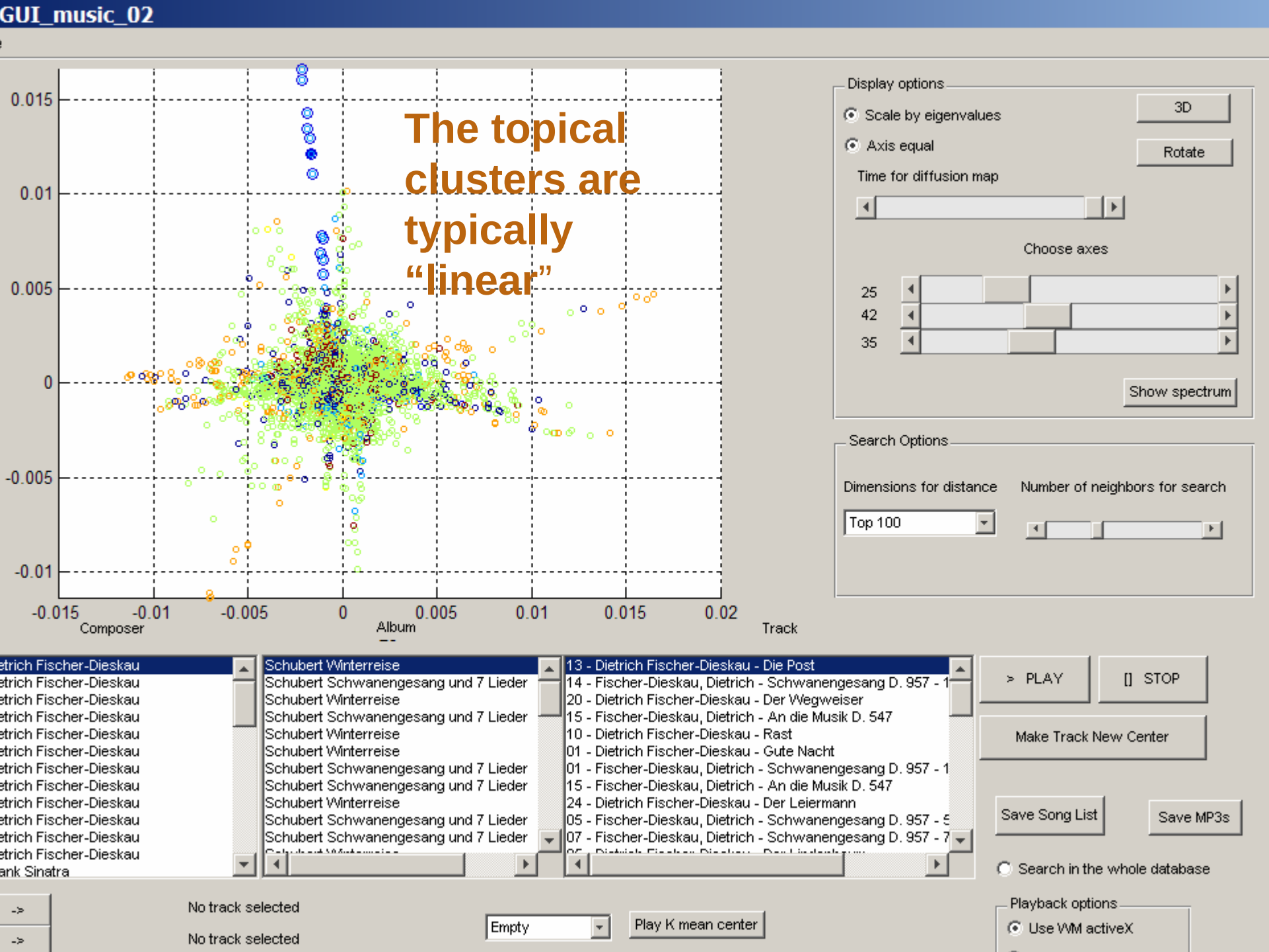
No track selected

Empty

Play K mean center

Playback options

Use WM activeX


GUI_music_02
The topical clusters are typically "linear"
0.015
0.01
0.005
0
-0.005
-0.01
-0.015
-0.01
-0.005
0
0.005
0.01
0.015
0.02
Composer
Album
Track
Display options
Scale by eigenvalues
Axis equal
3D
Rotate
Time for diffusion map
Choose axes
25
42
35
Show spectrum
Search Options
Dimensions for distance
Number of neighbors for search
Top 100
etrich Fischer-Dieskau
ank Sinatra
Schubert Winterreise
Schubert Schwanengesang und 7 Lieder
13 - Dietrich Fischer-Dieskau - Die Post
14 - Fischer-Dieskau, Dietrich - Schwanengesang D. 957 - 1
20 - Dietrich Fischer-Dieskau - Der Wegweiser
15 - Fischer-Dieskau, Dietrich - An die Musik D. 547
10 - Dietrich Fischer-Dieskau - Rast
01 - Dietrich Fischer-Dieskau - Gute Nacht
01 - Fischer-Dieskau, Dietrich - Schwanengesang D. 957 - 1
15 - Fischer-Dieskau, Dietrich - An die Musik D. 547
24 - Dietrich Fischer-Dieskau - Der Leiermann
05 - Fischer-Dieskau, Dietrich - Schwanengesang D. 957 - 5
07 - Fischer-Dieskau, Dietrich - Schwanengesang D. 957 - 7
> PLAY
[] STOP
Make Track New Center
Save Song List
Save MP3s
Search in the whole database
->
->
No track selected
No track selected
Empty
Play K mean center
Playback options
Use WM activeX