MMDS- Stanford 2008

Harmonic Analysis, diffusion geometries and Multi Scale organizations of data and matrices .

R.R Coifman

Department of 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,V.Rokhlin, S.W. Zucker.

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

Inferential/diffusion geometries on digital data graphs, enable the organization and analysis of empirical data as well as "signal processing" of functions on data.

In particular we will describe various natural multiscale structures on data which enable automatic ontology and "language building" for abstract digital data.

These developments extend geometries of spectral graph theory, kernel machines and other machine learning tools.

Digital data clouds can be organized through an affinity kernel A(x,y) where expert knowledge enters to build associations between documents. Such affinity is only robust for "nearest neighbors".

Two basic approaches for organizing data

I. ***Hierarchical folder building and clustering , a bottom up approach which propagates or diffuses affinity between documents=points . Can be achieved through probabilistic model building and statistical/combinatorial "book keeping" on the data***

II. ***A dimensional reduction approach which embeds the data in low dimensional Euclidean space , through the use of eigenvectors of the affinity kernel A (or related Matrix) followed by clustering and processing in that dimension.***

These two approaches, seemingly different, can be shown to be mathematically equivalent through the introduction of multiscale "inferential folder" structures based on affinity diffusions.

The eigenvectors are global functions on the data which "integrate" precisely the local "infinitesimal" affinity geometry.

Conversely "wavelet like" functions defined by affinity folders enable efficient "informed embeddings" of the data in low dimensional spaces. (As well as an efficient synthesis of the eigenfunctions .)

Overview

- ***Eigenvector "magic".***
- ***Diffusion geometry , eigenvectors as an extension of Newtonian calculus.***
- ***Multiscale geometry , localization of eigenvectors***

folder geometries or automatic ontologies. (the Zygmund program)

- ***The analysis of operators ,or data matrices such as questionnaires .***

We now illustrate the relation to multiscale organization

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).

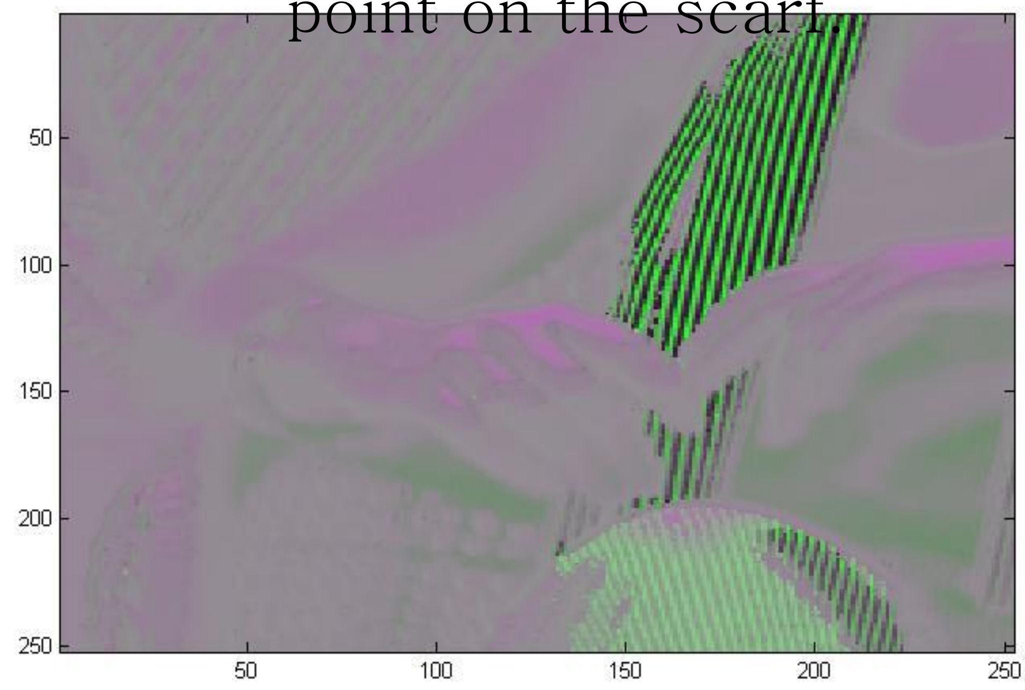
We illustrate the role of graph harmonic analysis to process complex data such as images .

Given an image, associate with each pixel p a vector v(p) of features . For example the 5x5 subimage centered at the pixel ,or any combination of features . Define a Markov matrix A as

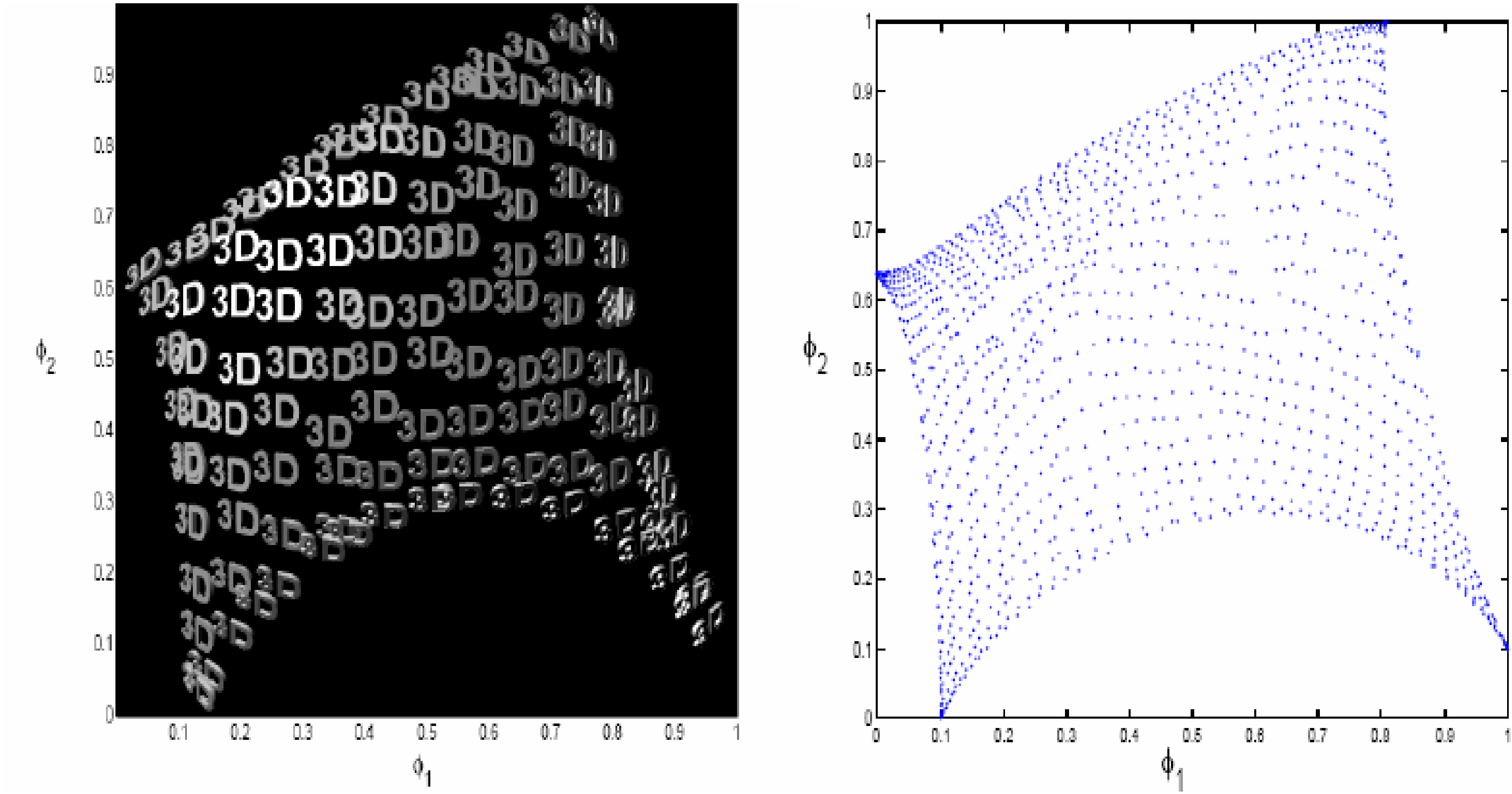
$$A_{p,q} = \frac{\exp(-\|v(p) - v(q)\|^2 / \varepsilon)}{\sum_q \exp(-\|v(p) - v(q)\|^2 / \varepsilon)}$$

We claim that the eigenvectors of this matrix contain all the geometric information concerning the structure of the image

The image on the left is projected into the three dimensional space spanned by the eigenvectors 5 ,8 10 (red, green, blue) which are active on a chosen point on the scarf.

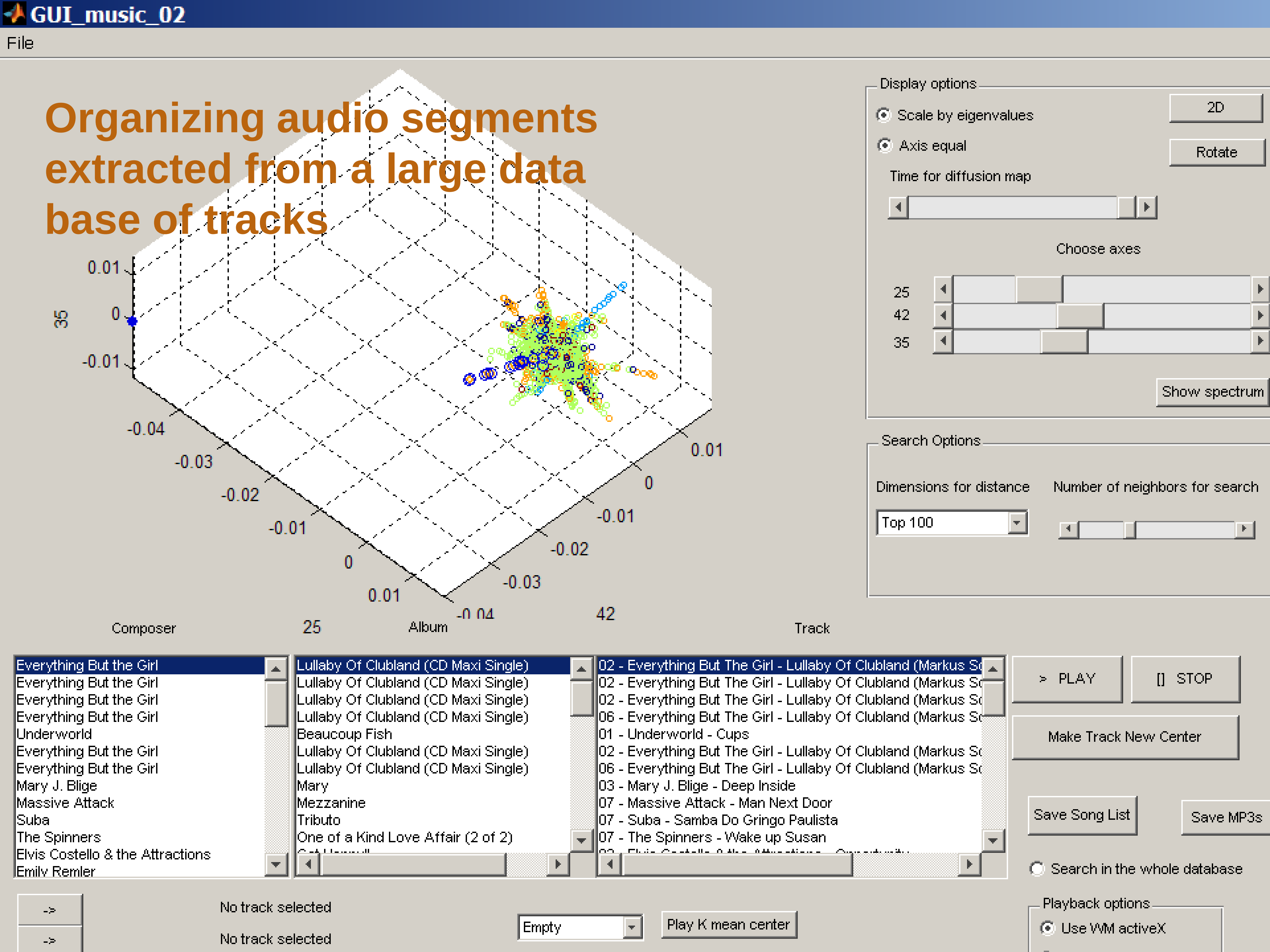


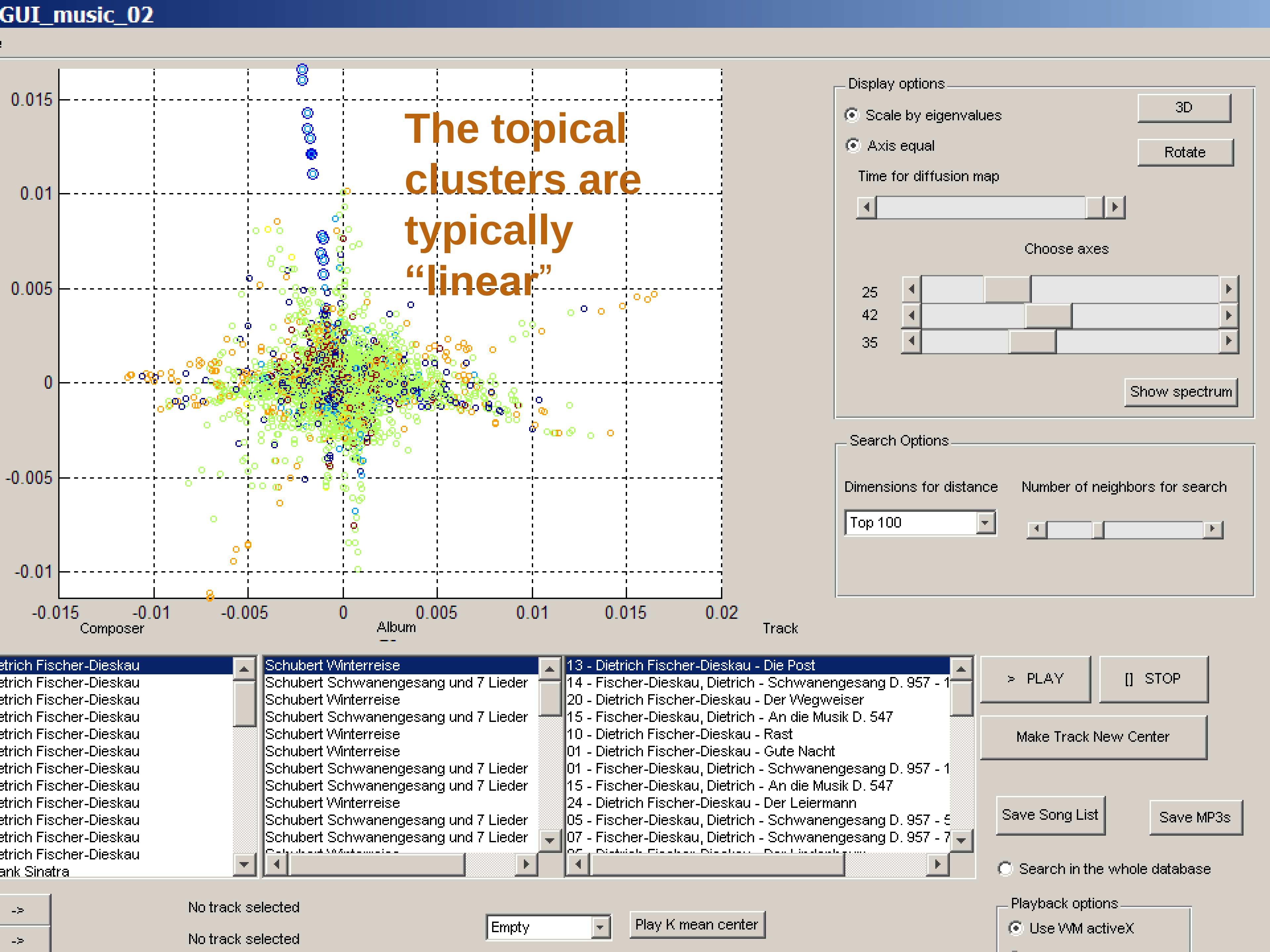
The image above is viewed as a data base of all sub images of size 5x5, natural structures are discovered through projections on various



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

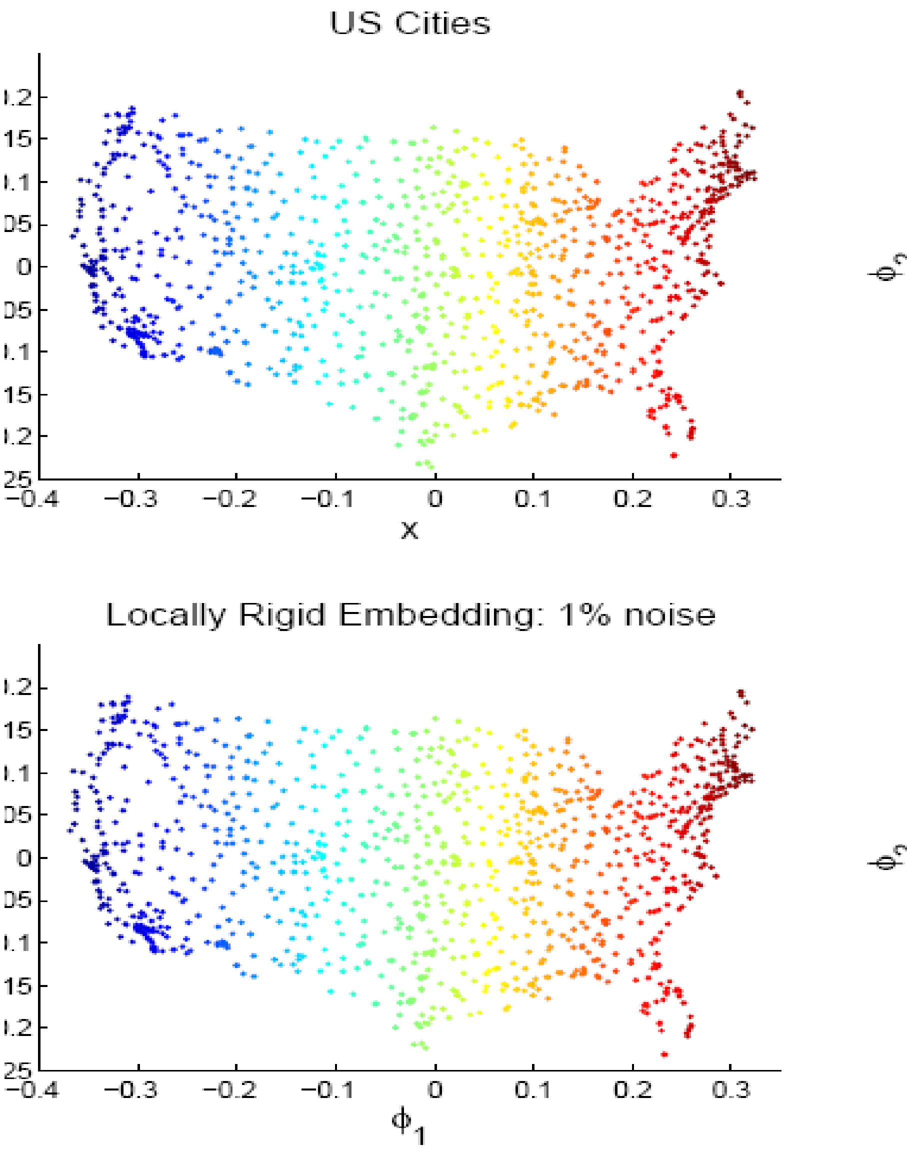
Organizing audio segments extracted from a large data base of tracks




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
Playback options
Use WM activeX
->
->
No track selected
No track selected
Empty
Play K mean center

A simple way to understand the relation between eigenvectors and geometry is provided by ***The sensor network local to global positioning problem***.

For each city in the US we know the distance to a few neighbors , how do we get the global position ?



Solution by A. Singer   may 2007

Let P_i be the location of city (or sensor) i . From the knowledge of the distance to a few neighbors P_j we can easily calculate from local connections weights $W_{i,j}$ so that :

$$P_i = \sum_{j \neq i} W_{i,j} P_j \quad \text{where} \quad \sum_{j \neq i} W_{i,j} = 1$$

Clearly both x and y coordinates (as well as 1) are eigenvetors of the matrix W.

The matrix W is a local encapsulation of the relation between cities.

Generalization of the fundamental theorem of calculus .

Assume that at each site you know the difference of altitude between cities and some of their neighbours we get the global function as the z eigenfunction of the 3 dimensional version .

Basically find the altitude function from its local increments.

We observe that given f we can easily solve the Poisson equation (or any other "differential equation") on graphs

$\Delta u = f$, where $\Delta = I - A$,

and A is any local averaging operator.

In fact let $B = A + \alpha A \sigma (I - A)$ with $\alpha = \frac{f}{A|f|}, \sigma = \mathrm{sgn}(f)$,

It is easy to check that

$$Bu = u$$

and therefore the solution to the Poisson equation is an eigenvector of B with eigenvalue 1.

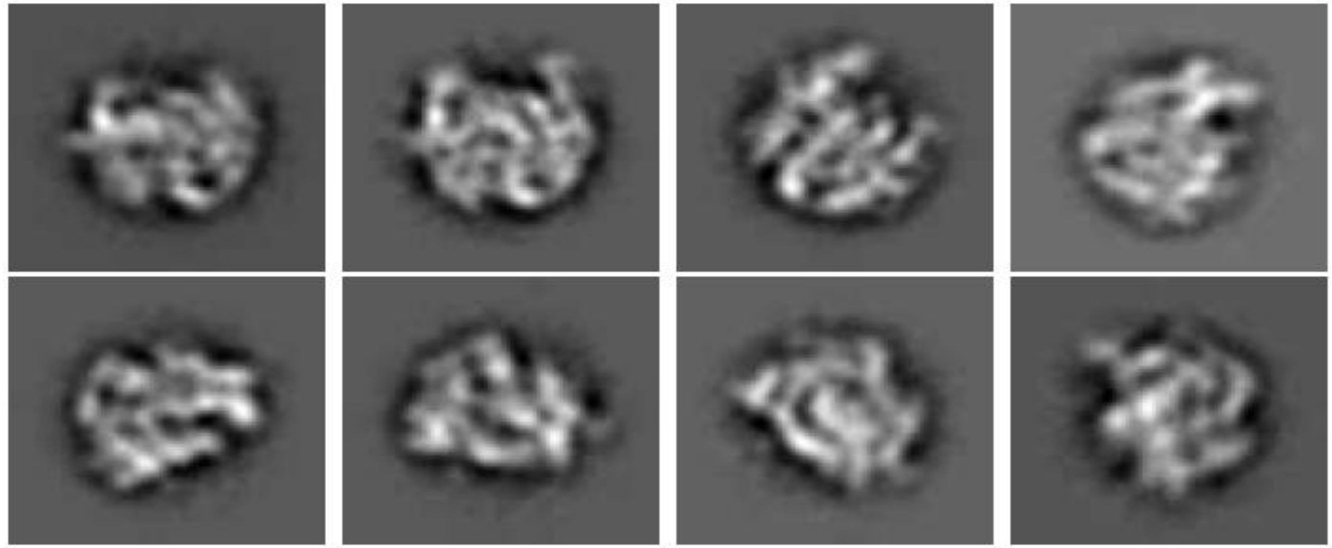
Cryo-Microscopy

Application

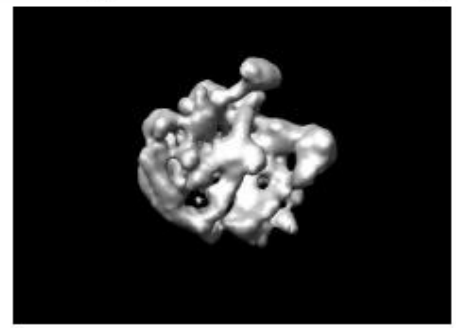
Example of E Coli

Observed from random

angles .



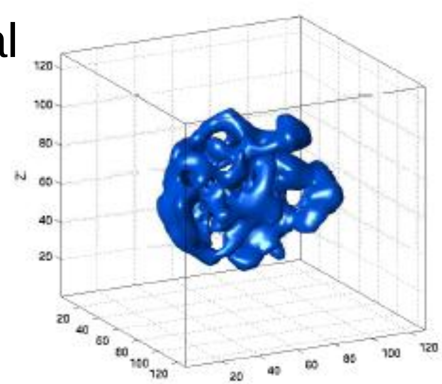
The full three d picture is rebuild from knowledge of local angular distances using the center of mass method

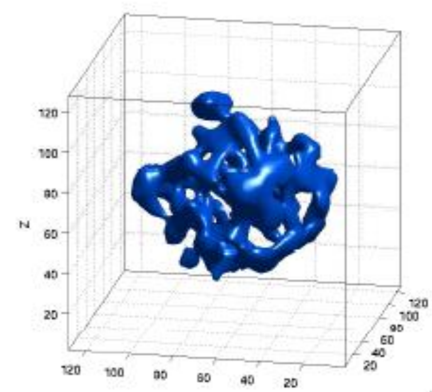


(a) Original

A similar protein reconstruction from NMR enables to rebuild in a few seconds on a laptop a structure that currently takes hours on a supercomputer using conventional optimization.

reconstructions





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

Let X_i ***represent normalized data (they are simply rows of a data matrix) ,we "soft truncate" the covariance matrix defining an infinitesimal affinity 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 a discrete Graph Laplace Operator.

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

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

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

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

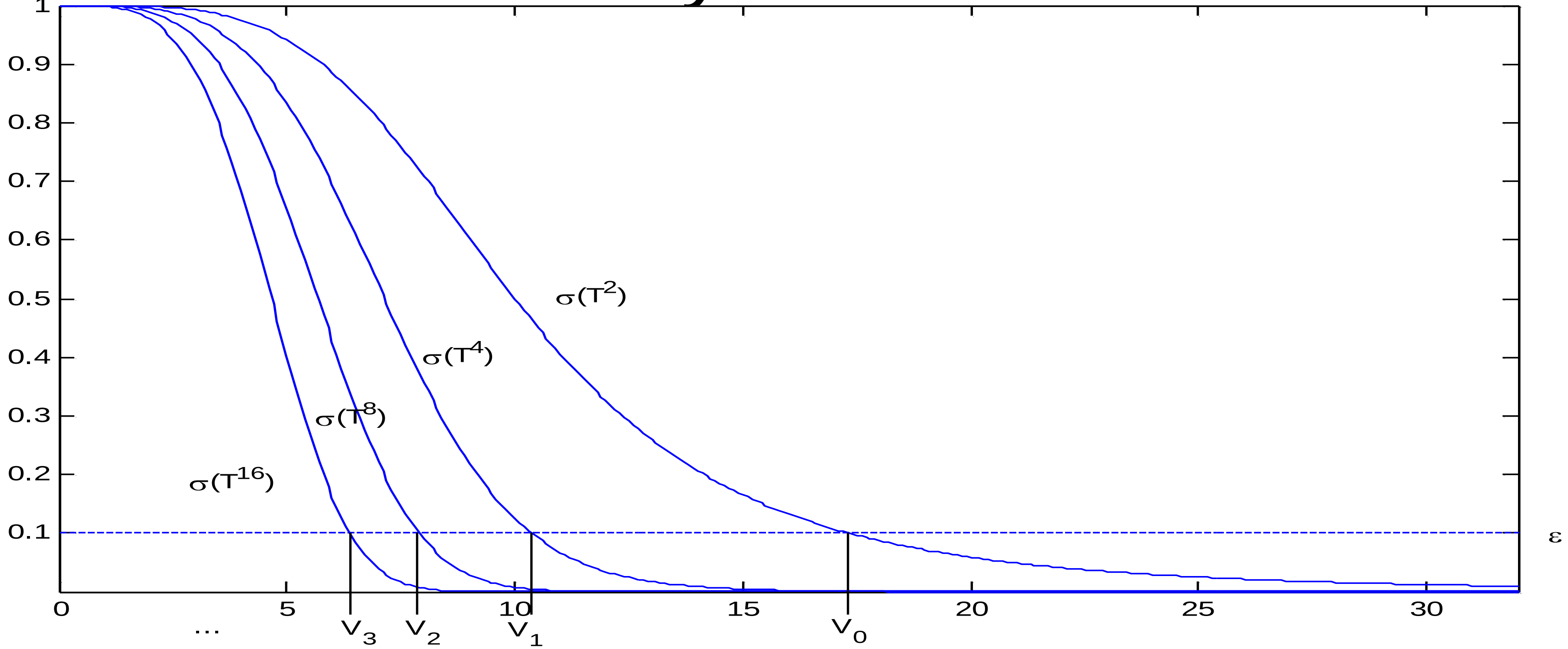
An alternative affinity matrix between points

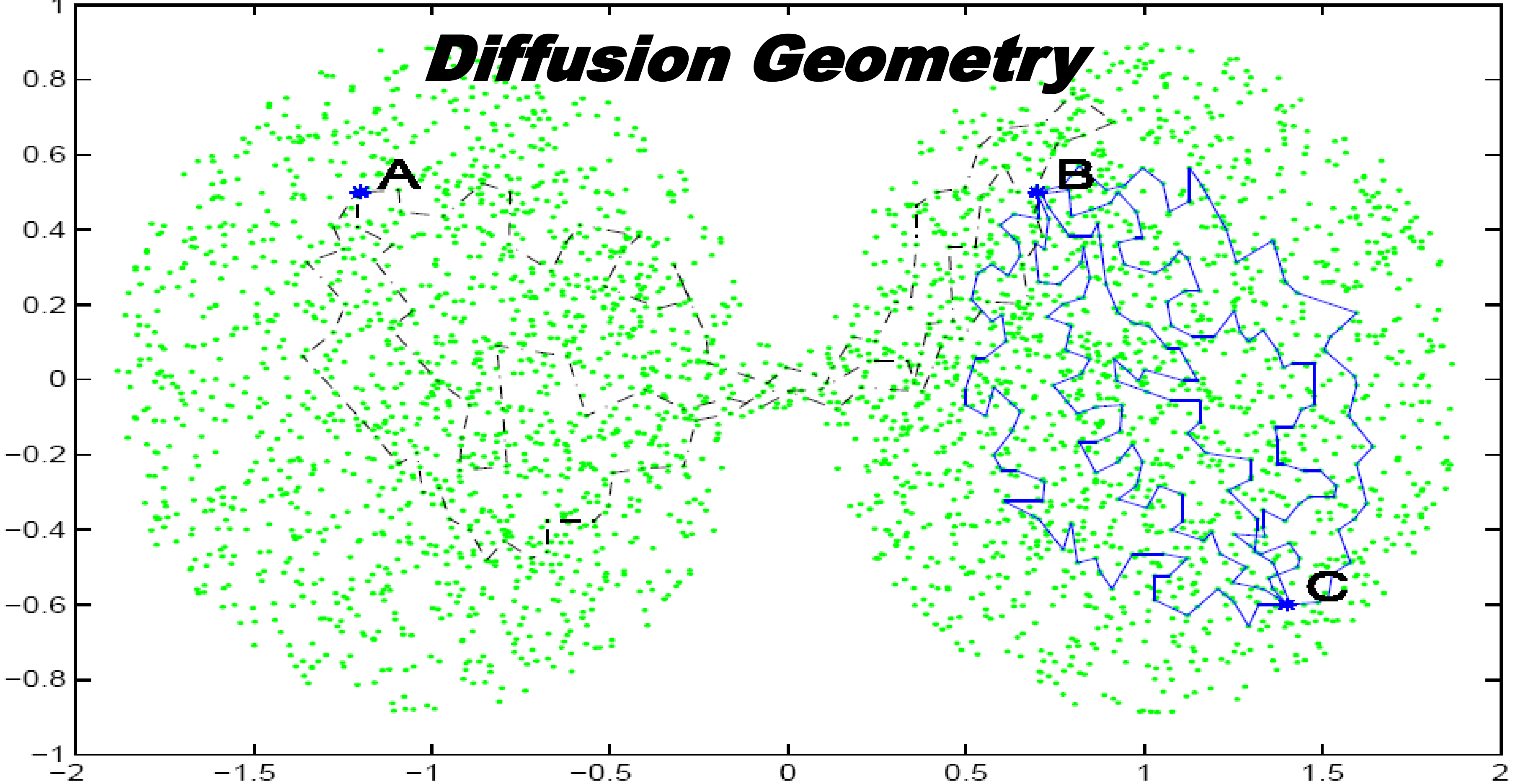
$$A_{i,j} = \frac{\exp(-\|x_i - x_j\|^2 / \varepsilon)}{\omega_i \omega_j}$$

Where the weights are selected so that A is Markov or similar to a Markov matrix defining a diffusion on the cloud of points x_i .

If we consider the spectrum of the various powers of the diffusion operator A we see that its numerical rank can drop dramatically.

This property enables both data filtering and multiscale analysis

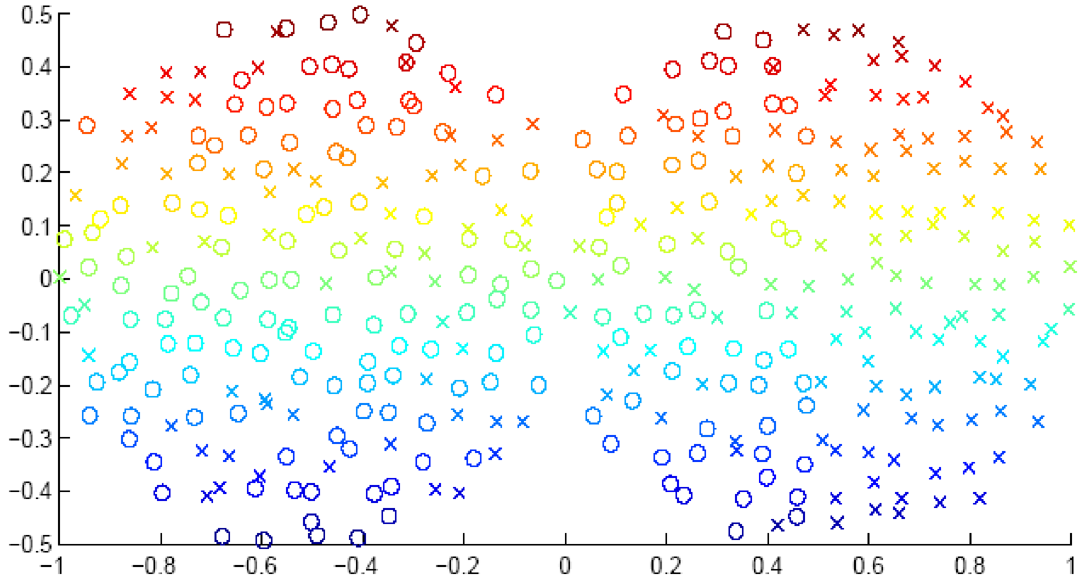


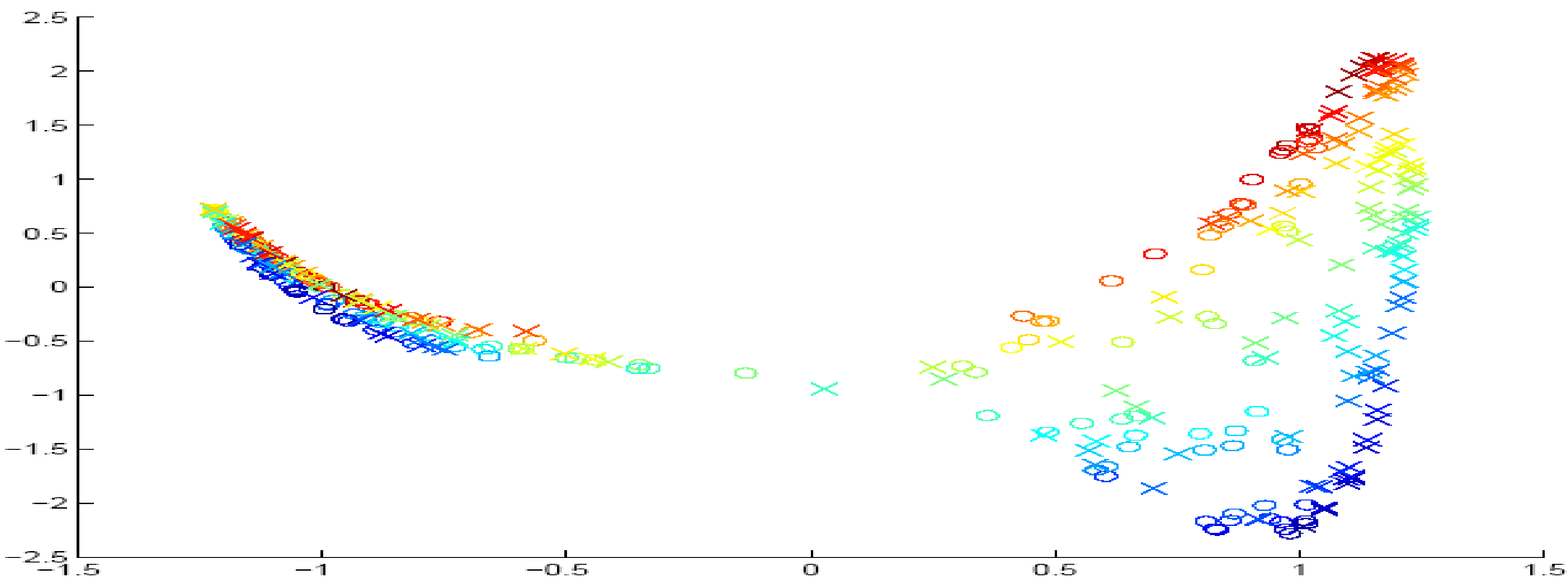


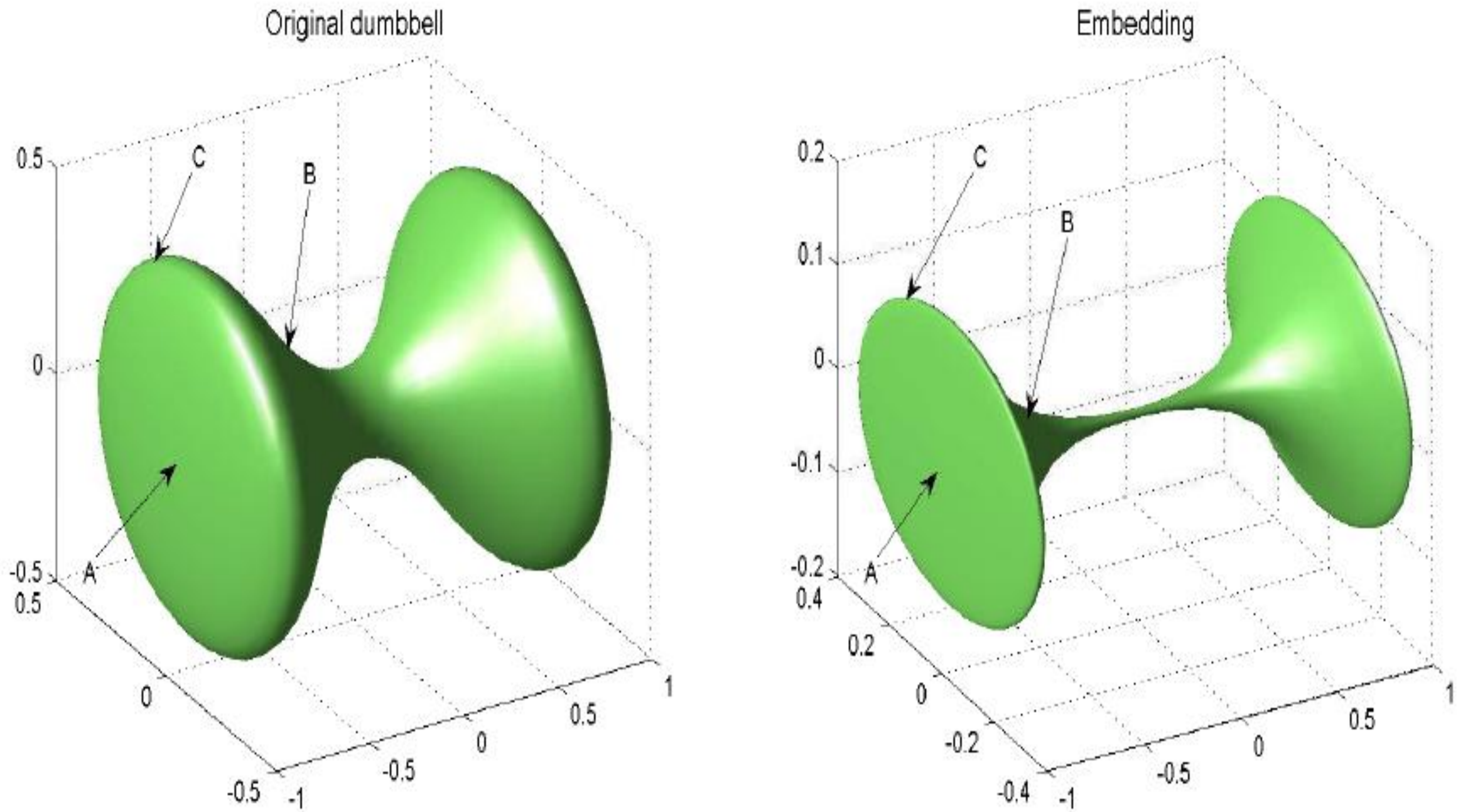
Diffusions between A and B have to go through the bottleneck ,while C is easily reachable from B. The Markov matrix defining a diffusion could be given by a kernel , or by inference between neighboring nodes.

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.

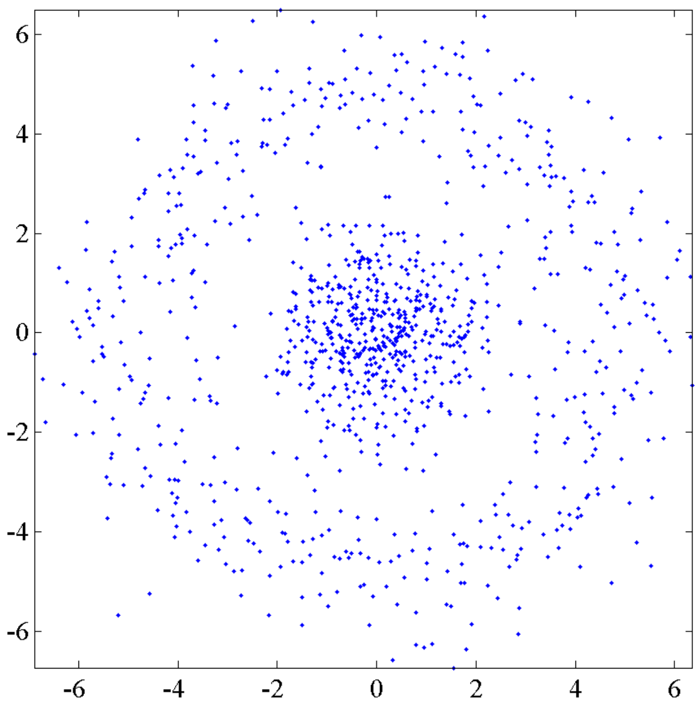
The long term diffusion of heterogeneous material is remapped below . The left side has a higher proportion of heat conducting material ,thereby reducing the diffusion distance among points , the bottle neck increases that distance



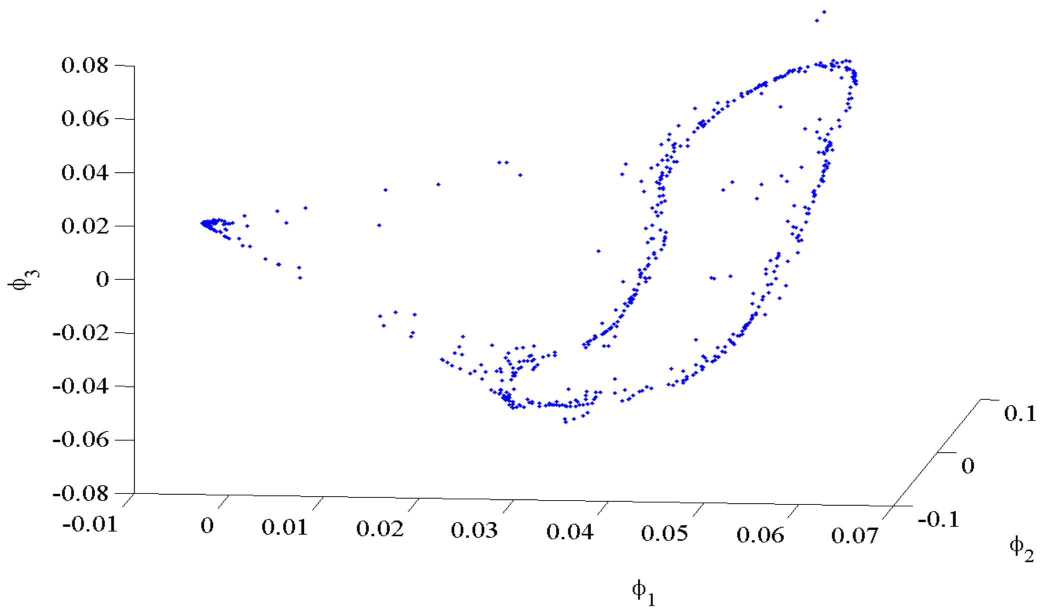




The natural diffusion on the surface of the dumbbell is mapped out in the embedding . Observe that A is closer to B than to C ,and that the two lobes are well separated by the bottleneck.



Original data set

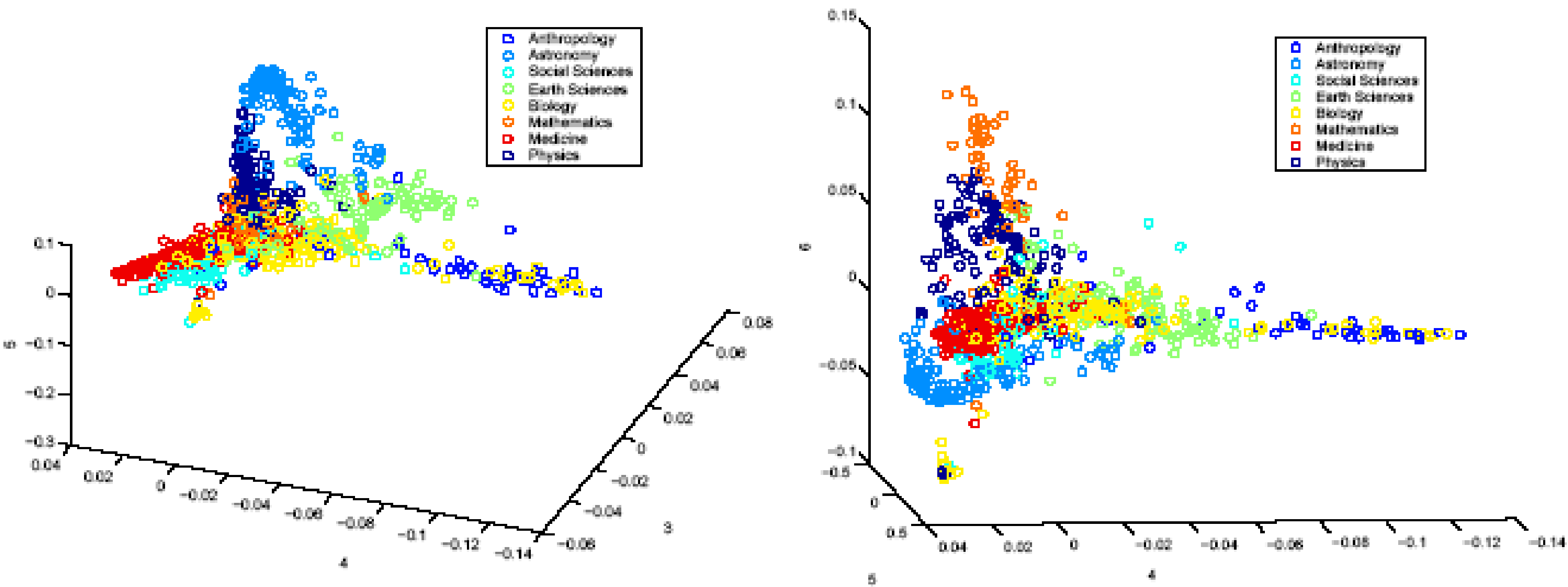


Embedding of data into the first 3 diffusion coordinates

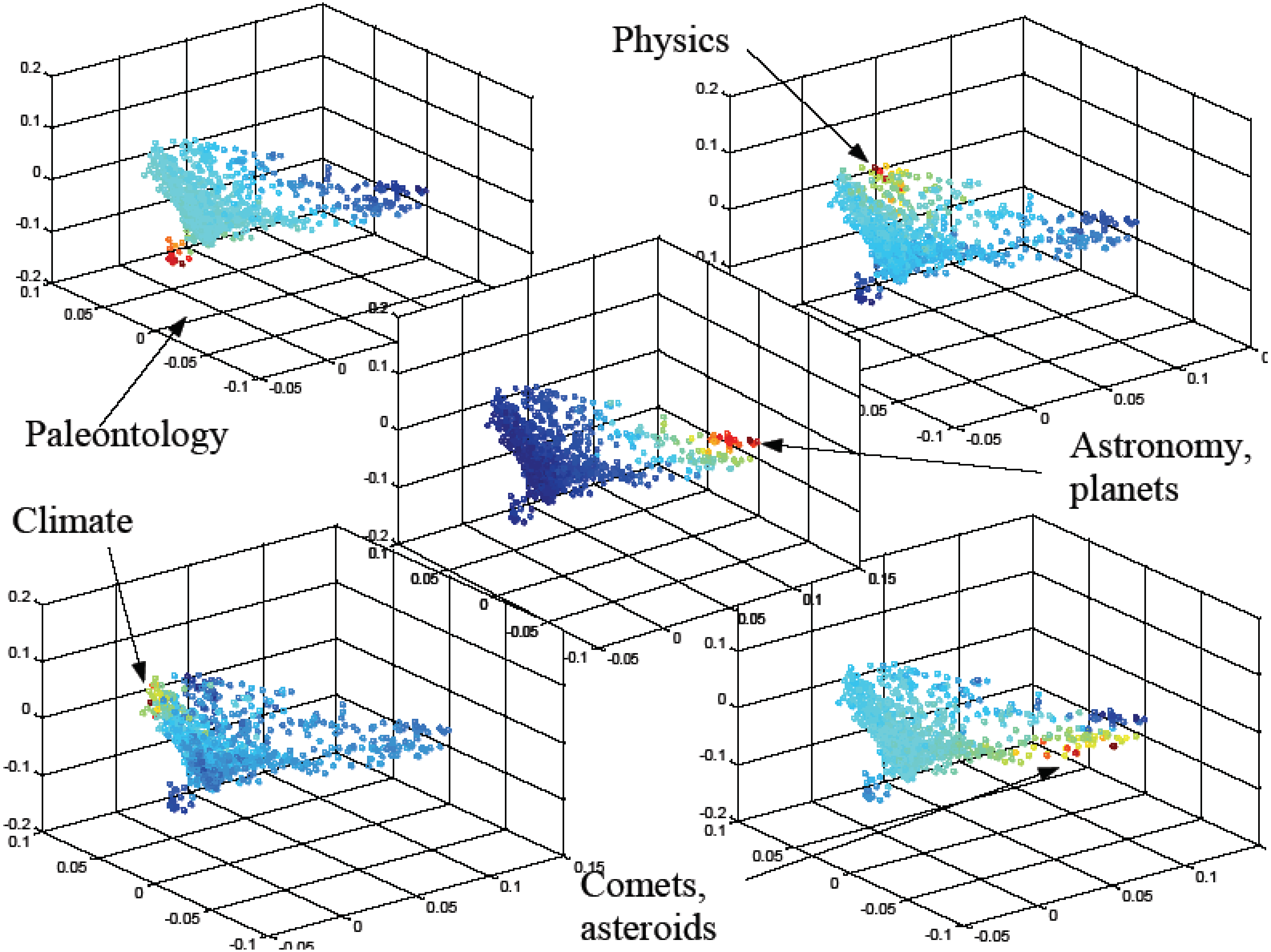
Here organization is achieved through ,eigenfunctions and wavelet constructions

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.



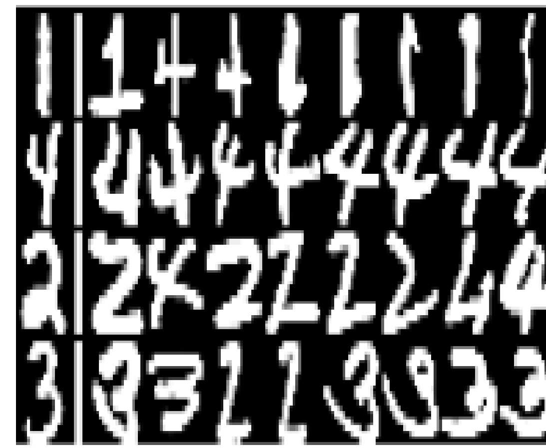
Embedding $\Xi_6^{(0)}(x) = (\xi_1(x), \ldots, \xi_6(x))$: on the left coordinates $3, 4, 5$, and on the right coordinates $4, 5, 6$.

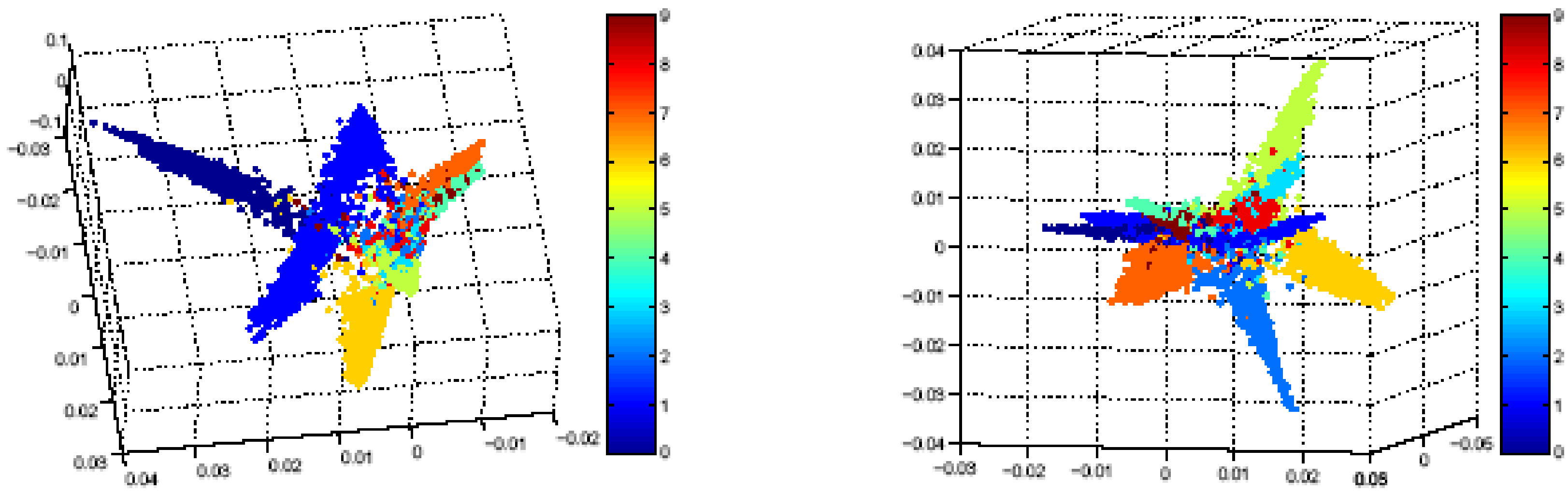


Physics
Paleontology
Astronomy, planets
Climate
Comets, asteroids

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 book keeping 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

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

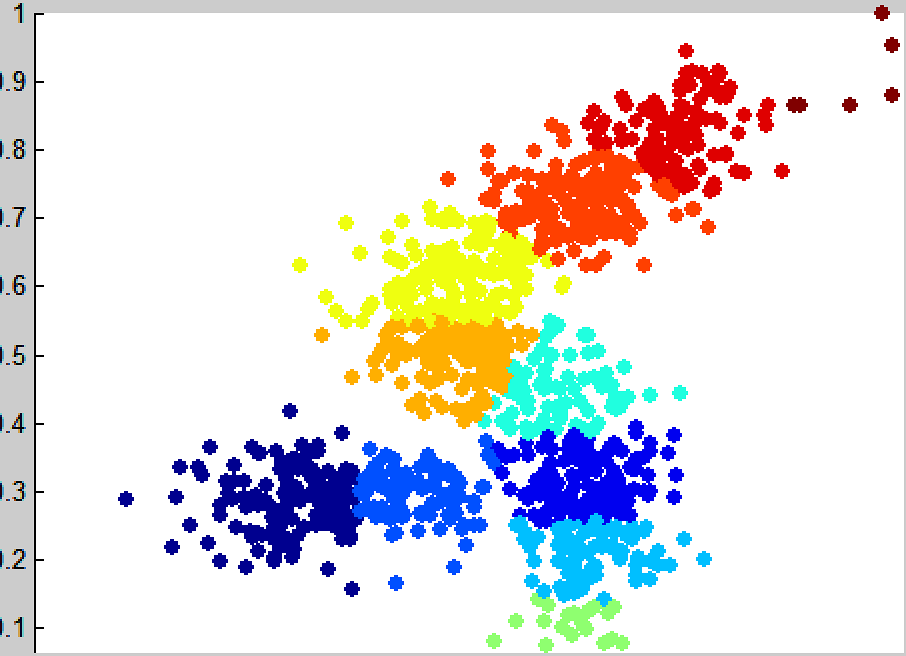
$$d_t^2(x,y) = a_t(x,x) + a_t(y,y) - 2a_t(x,y)$$

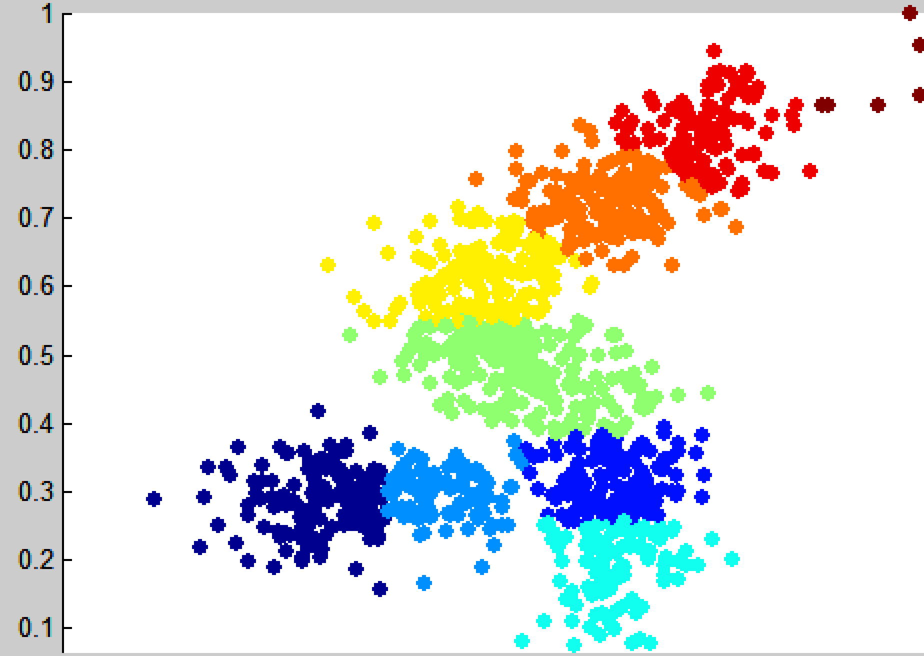
is the distance at scale t between x and y ,

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

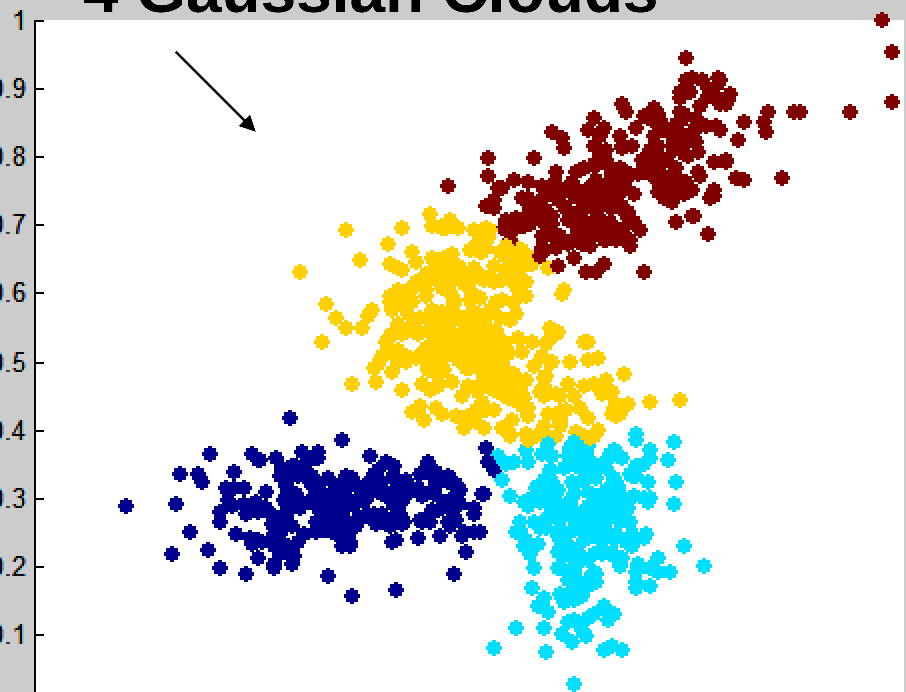
Start with a disjoint partition of the graph into clusters of diameter between 1 and 2 relative in the diffusion distance with t=2.
Consider the new graph formed by letting the elements of the partition be the vertices .
Using the distance between sets and affinity between sets described above we repeat 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)

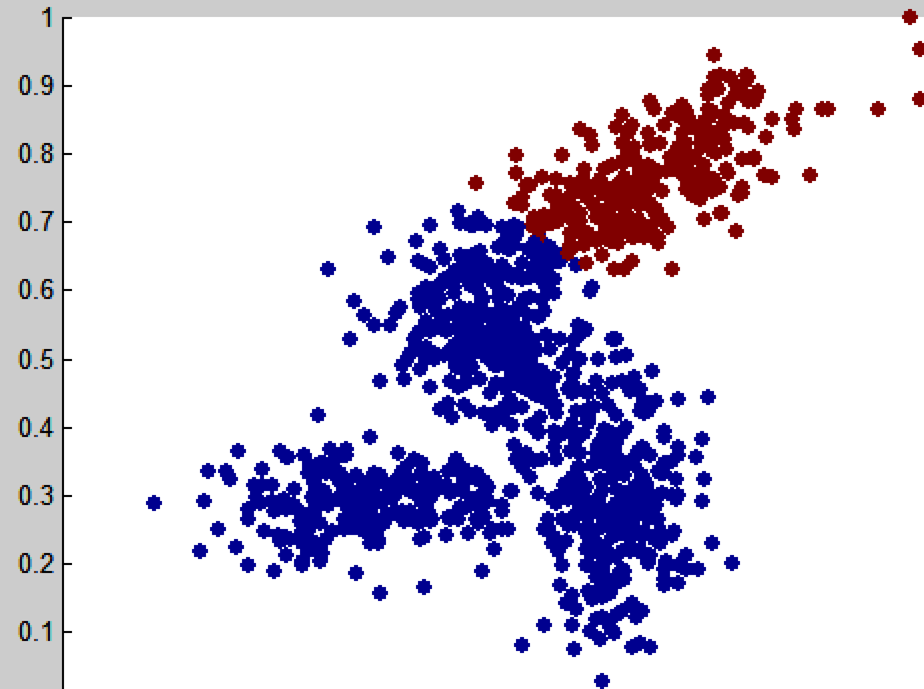
In the next image we see this organization as it applies to a random collection of 4 Gaussian clouds .



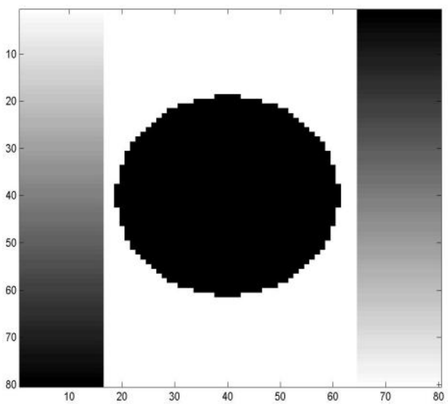


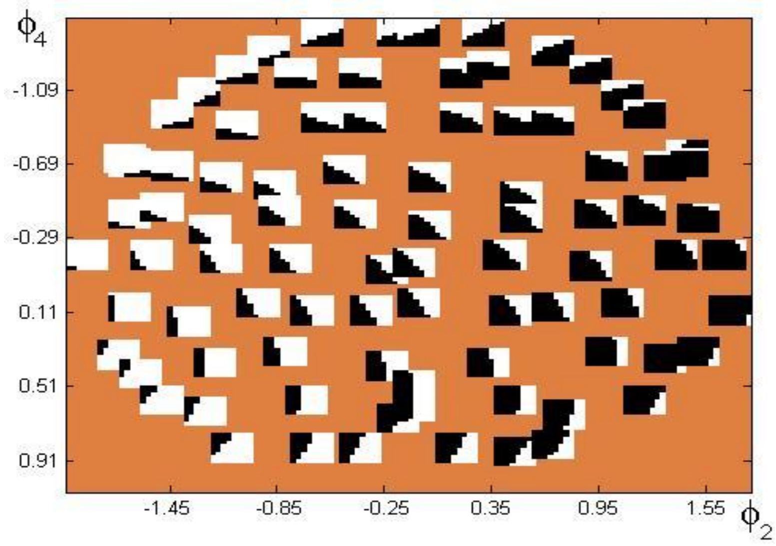
4 Gaussian Clouds

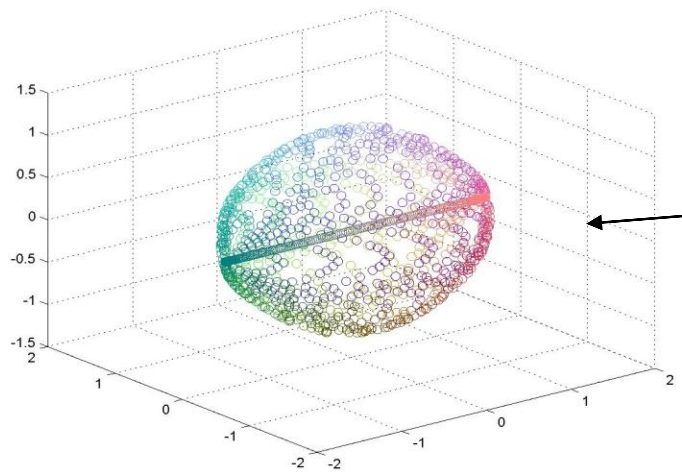




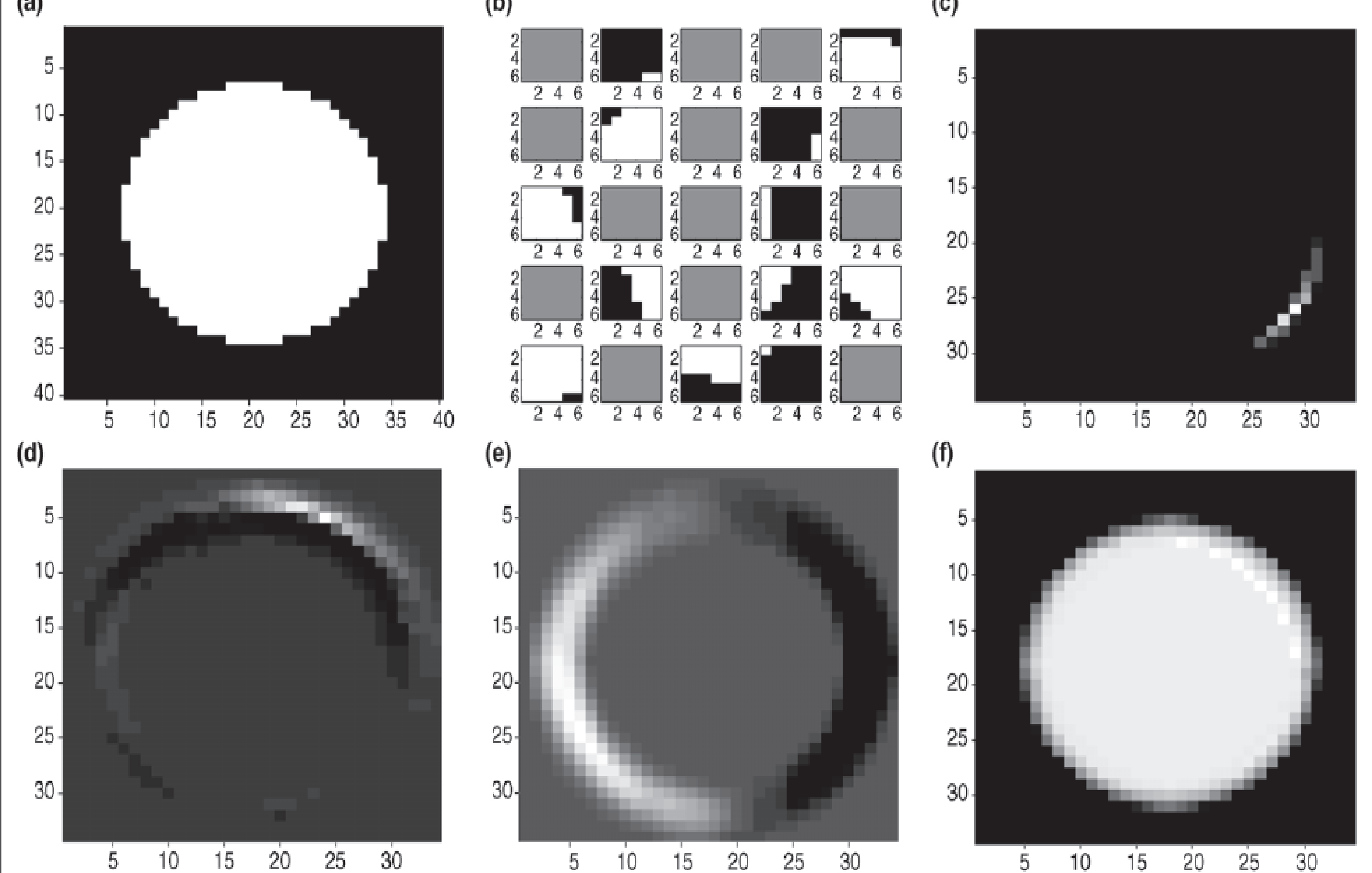
We now organize the set of subimages of 8x8 squares extracted from the left image and organized “naturally” by their average and orientation of the edge (the first two eigenfunction coordinates) .







The first 3 eigenfunctions describe the full geometry of this data .



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

We described a calculus of digital data as a first step in addressing and setting up many of the issues mentioned above ,and much more, including multidimensional document rankings extending Google, information navigation, heterogeneous material modeling, multiscale complex structure organization etc.

Remarkably this can be achieved with algorithms which **scale linearly with the number of samples.**
*The methods described below are also known as **nonlinear principal component analysis, kernel methods, support vector machines, spectral graph theory,** and many more They are documented in literally hundreds of papers in various communities.*
A simple description of many of these ideas and more is given through diffusion geometries. (see the July 2006 issue of Applied and Computational Harmonic Analysis).