Diffusion/Inference geometries of data features, situational awareness and visualization

Ronald R Coifman
Mathematics
Yale University

• Digital data is generally converted to point clouds in high dimensional Euclidean space .

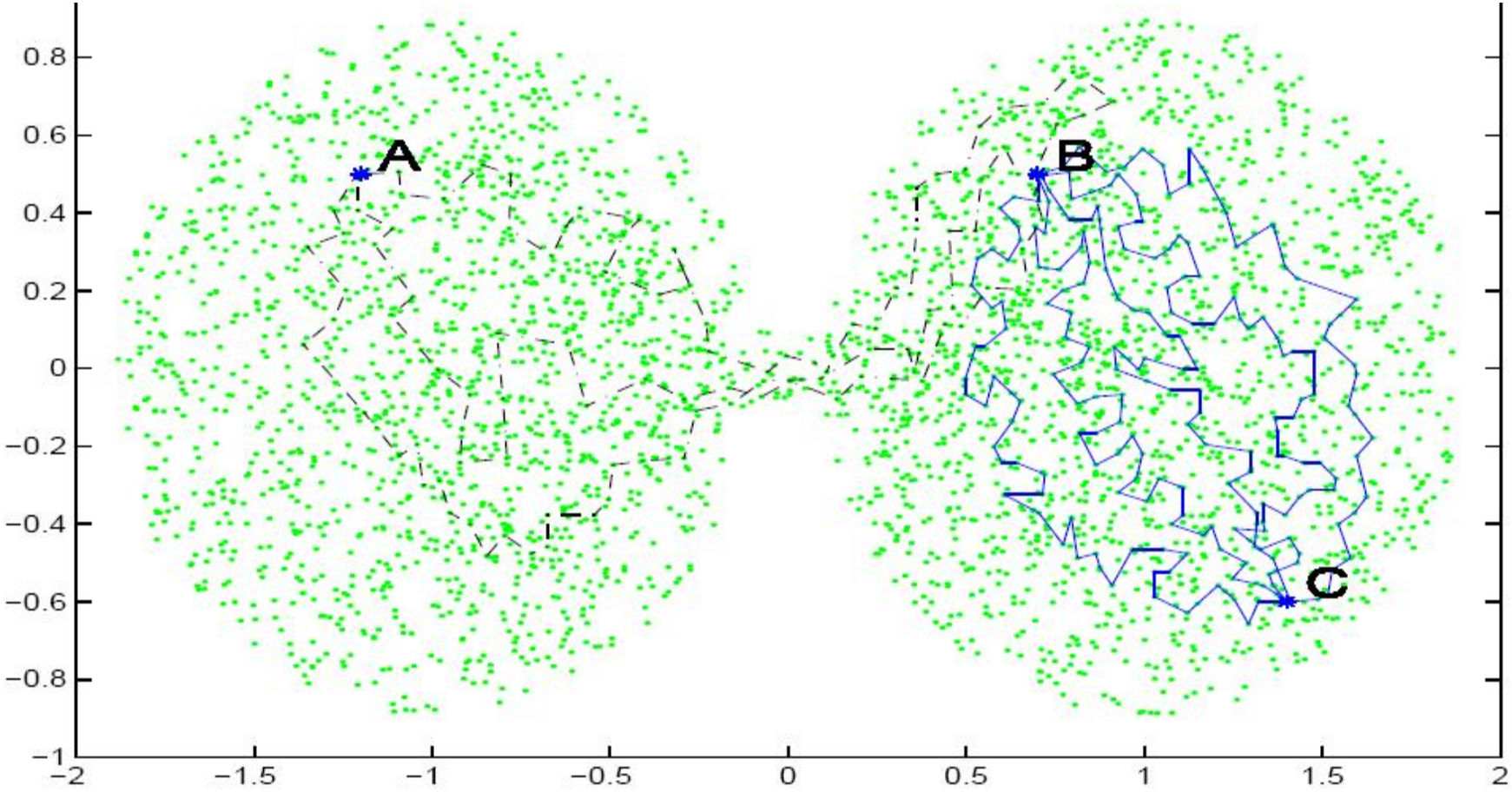
• The data is often correlated and analyzed through global linear algebra tools such as SVD.

• More generally a kernel defining affinity between data points can replace the correlation .(usually the kernel of a positive operator)

• Distances between data points are used to build a graph for which the spectral theory of the Laplace operator can be used to provide data clustering .

• In all these cases the eigenfunctions of some linear operator are used to effect a dimensional reduction through projection into a low dimensional space .

• ***We claim that diffusion (inference) geometries provide a powerful systematic approach to data analysis and organization, a methodology which enables global situational assessment based on local inferences .***



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 . 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.

Diffusion Maps

Let $\{\boldsymbol{x}_i\}_{i=1}^N$ denote a set of N points in $\mathbb{R}^p$.

View collection of data as a graph with N vertices and with connection strength between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ given by $k_\varepsilon(\boldsymbol{x}_i, \boldsymbol{x}_j)$, where

$$k_\varepsilon(\boldsymbol{x}, \boldsymbol{y}) = \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{y}\|^2}{2\varepsilon}\right)$$

Construct a Markov chain random walk based on these weights:

$$M_{i,j} = \Pr\{\boldsymbol{x}(t+\varepsilon) = \boldsymbol{x}_i | \boldsymbol{x}(t) = \boldsymbol{x}_j\} = \frac{k_\varepsilon(\boldsymbol{x}_i, \boldsymbol{x}_j)}{p_\varepsilon(\boldsymbol{x}_j)}$$

where

$$p_\varepsilon(\boldsymbol{x}_j) = \sum_i k_\varepsilon(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

Claim: first few eigenvectors and eigenvalues of this matrix $\{\lambda_j, \phi_j\}$ contain useful geometric information.

Diffusion Maps / Normalized Graph Laplacian

The diffusion map at time t is defined as the non-linear embedding

$$\boldsymbol{x} \to \Phi_t(\boldsymbol{x}) = \left(\lambda_1^t \phi_1(\boldsymbol{x}), \lambda_2^t \phi_2(\boldsymbol{x}), ..., \lambda_k^t \phi_k(\boldsymbol{x})\right)$$

Diffusion Distances

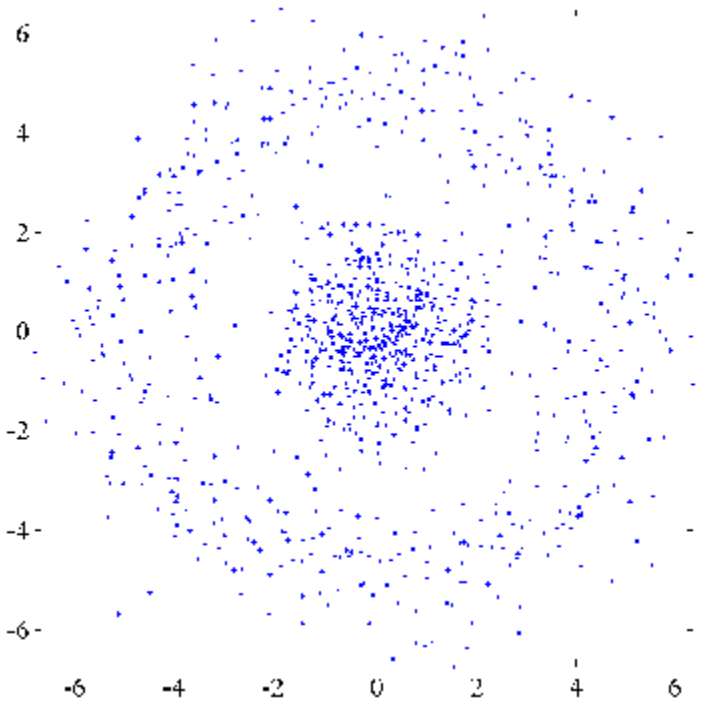
Definition: Diffusion distance at time t,

$$D_t^2(\boldsymbol{x}, \boldsymbol{y}) = M_s^t(\boldsymbol{x}, \boldsymbol{x}) + M_s^t(\boldsymbol{y}, \boldsymbol{y}) - 2M_s^t(\boldsymbol{x}, \boldsymbol{y})$$

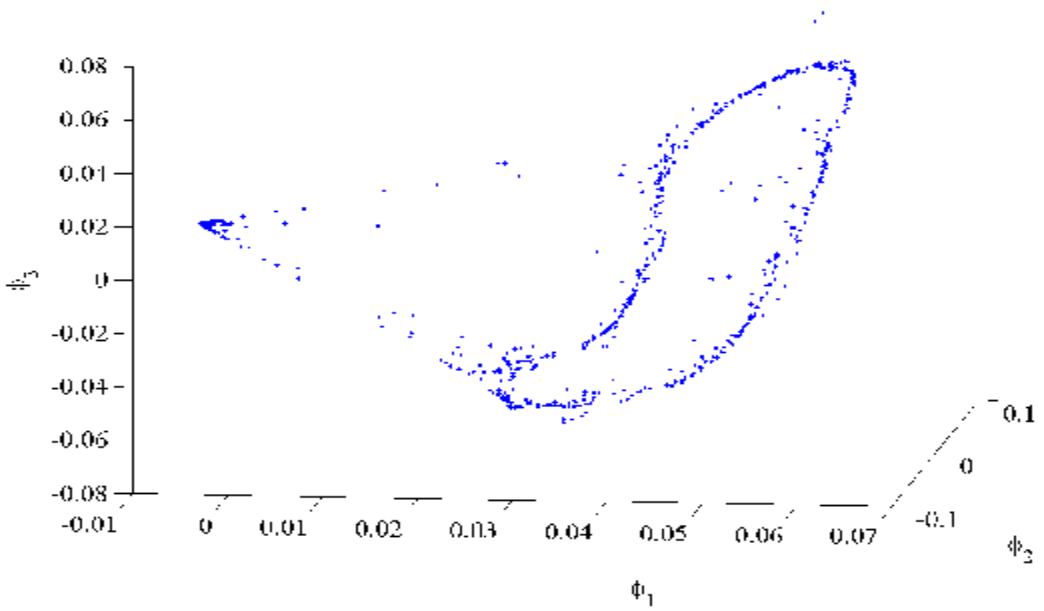
Spectral theorem:

$$D_t^2(\boldsymbol{x}, \boldsymbol{y}) = \sum_j \lambda_j^{2t} \left(\phi_j(\boldsymbol{x}) - \phi_j(\boldsymbol{y})\right)^2 = \|\Phi_t(\boldsymbol{x}) - \Phi_t(\boldsymbol{y})\|^2$$

Diffusion map converts the diffusion distance into Euclidian distance.

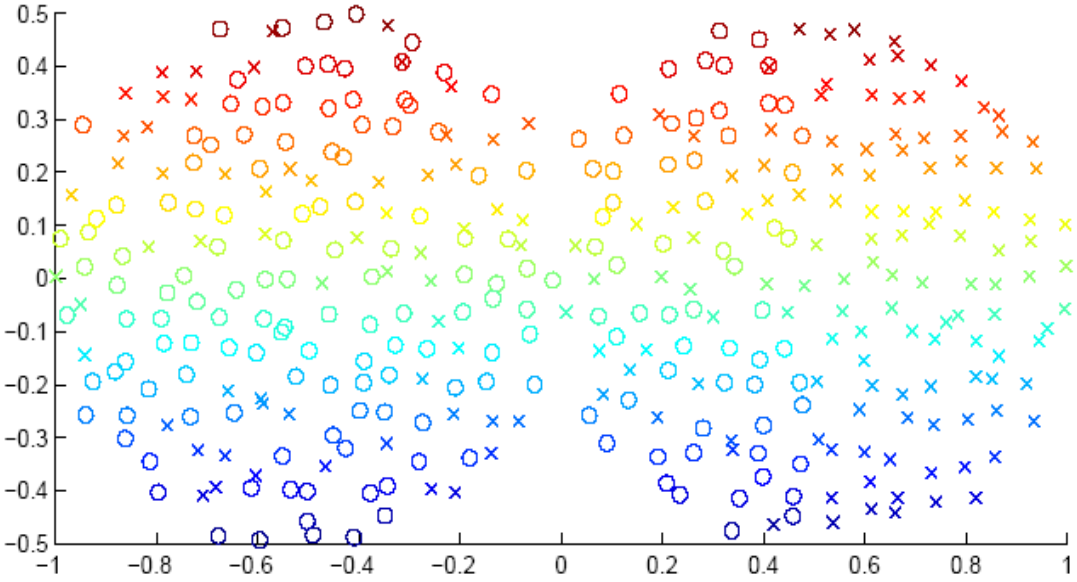


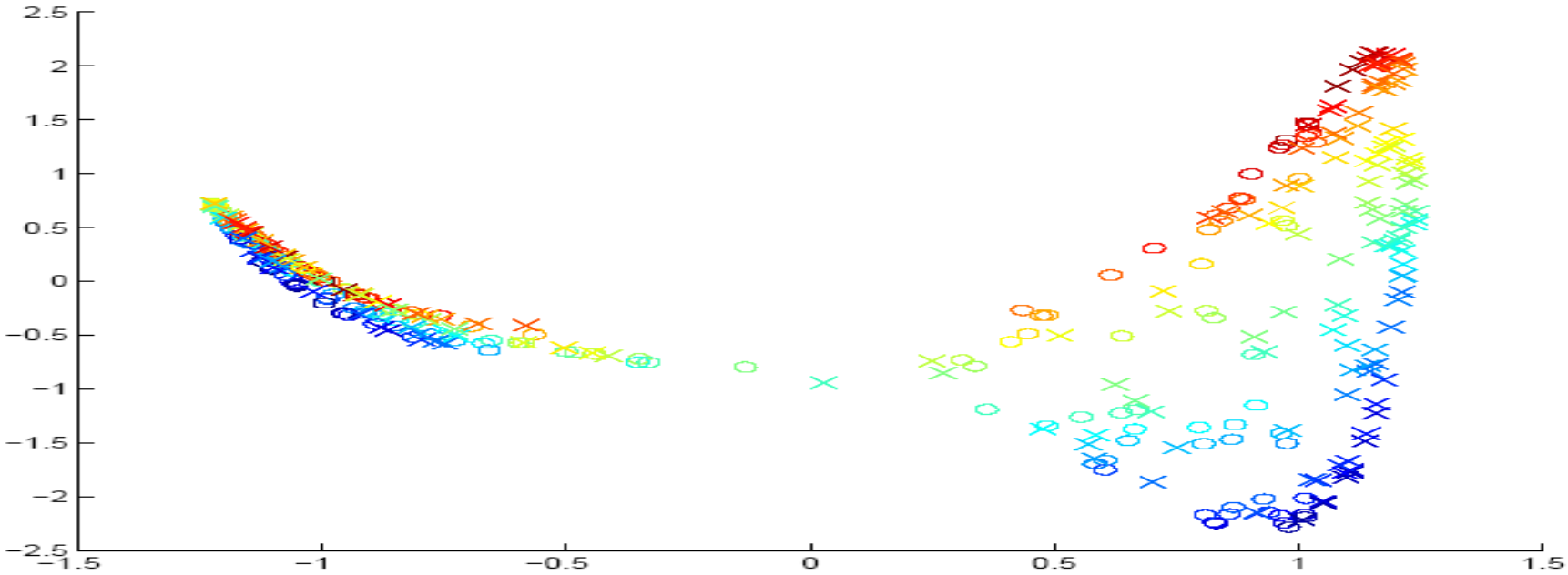
Original data set

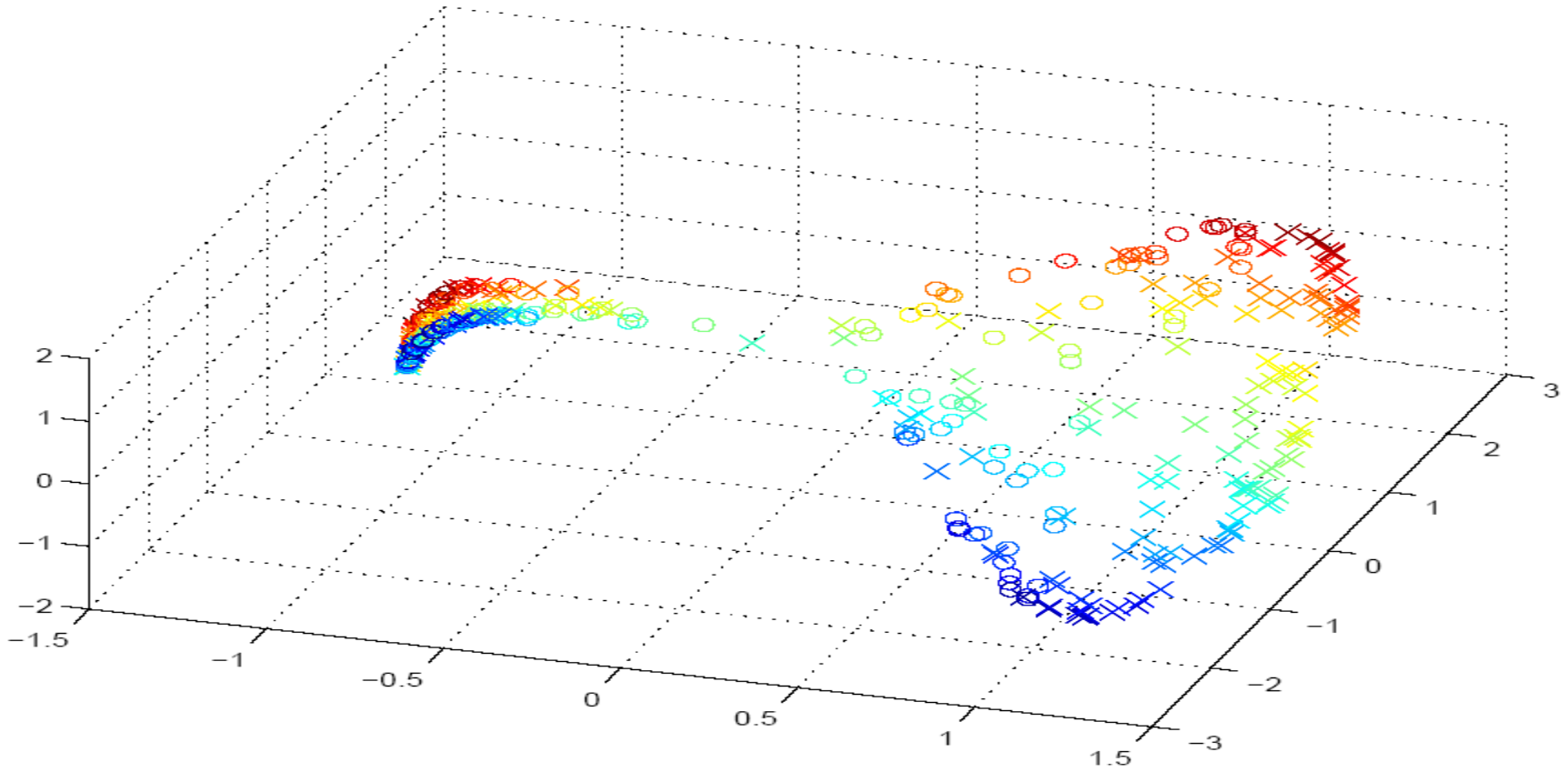


Embedding of data into the first 3 diffusion coordinates

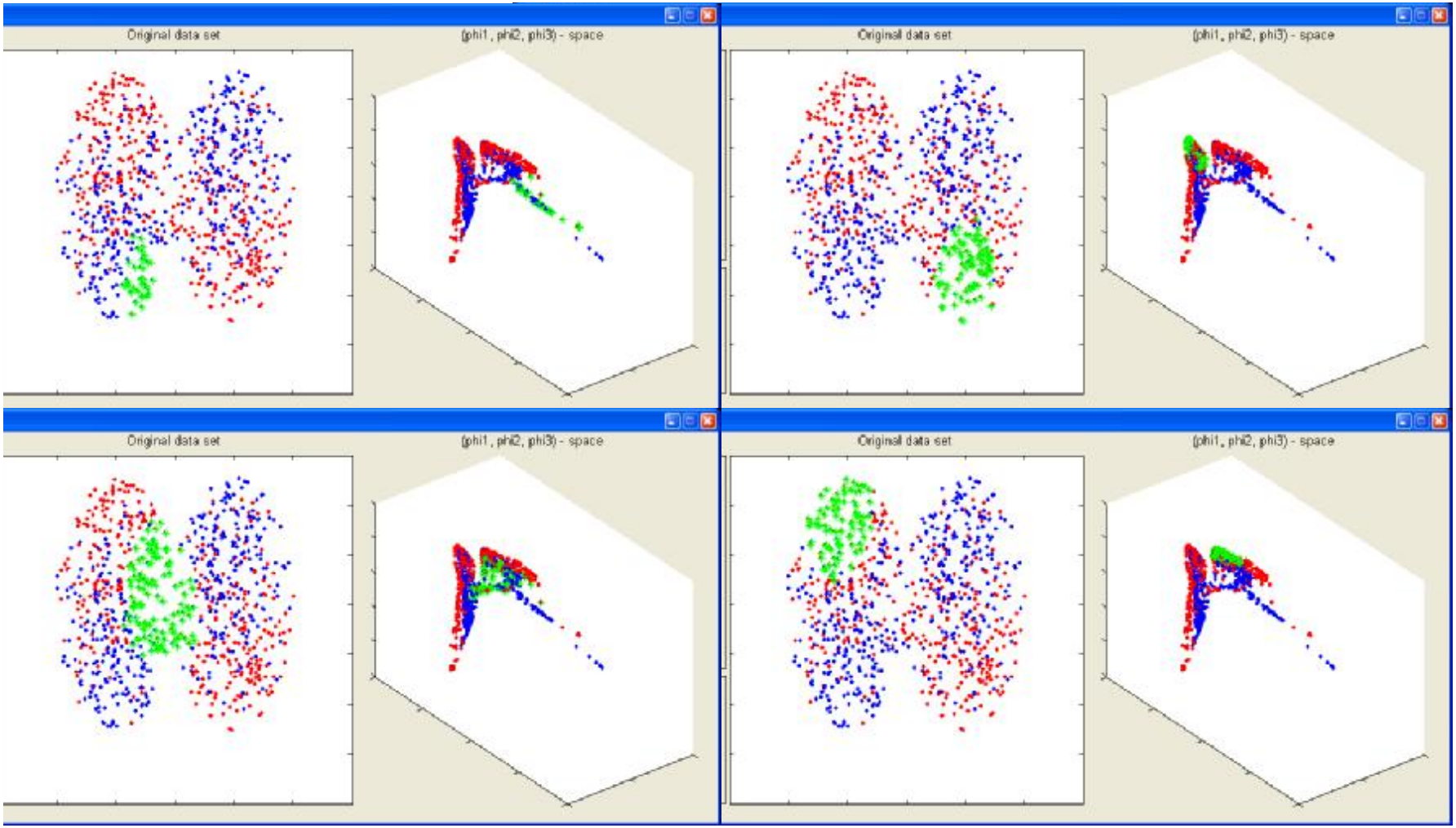
The long term diffusion of heterogeneous material is remapped below . The left side has a higher proportion of heat conducting material ,thereby reducing the diffusion distance among points , the bottle neck increases that distance







Diffusion map into 3 d of the heterogeneous graph
The distance between two points measures the diffusion between them.



A mixture of two materials where the red is very conductive relative to the blue The green sets are diffusion balls at time 3 .

On the right we have the diffusion map which encodes the diffusion distance as the Euclidean distance in the map (in order number of samples, not its square).

•*A similarity or relationship between two points (digital documents) is computed as a combination of* ***all chains of inference*** *linking them , these are the diffusion metrics .*

•*Clustering in this metric leads to* ***robust document segmentation*** *and tracking.*

•*Various local criteria of relevance lead to distinct geometries. In these geometries the* ***user can define relevance and filter away unrelated information.***

•***Self organization of digital documents*** *is achieved through local similarity modeling , the top eigenfunctions of the empirical model provide global organization of the given set of data.*

•*The diffusion maps embed the data into low dimensional Euclidean space and* ***convert isometrically the (diffusion) relational inference*** *metric to the corresponding Euclidean distance .*

•*Diffusion metrics can be* ***computed efficiently*** *as an ordinary Euclidean distance in a low dimensional embedding by the diffusion map (total computation time scale linearly with the data size, and can be updated on line) .*

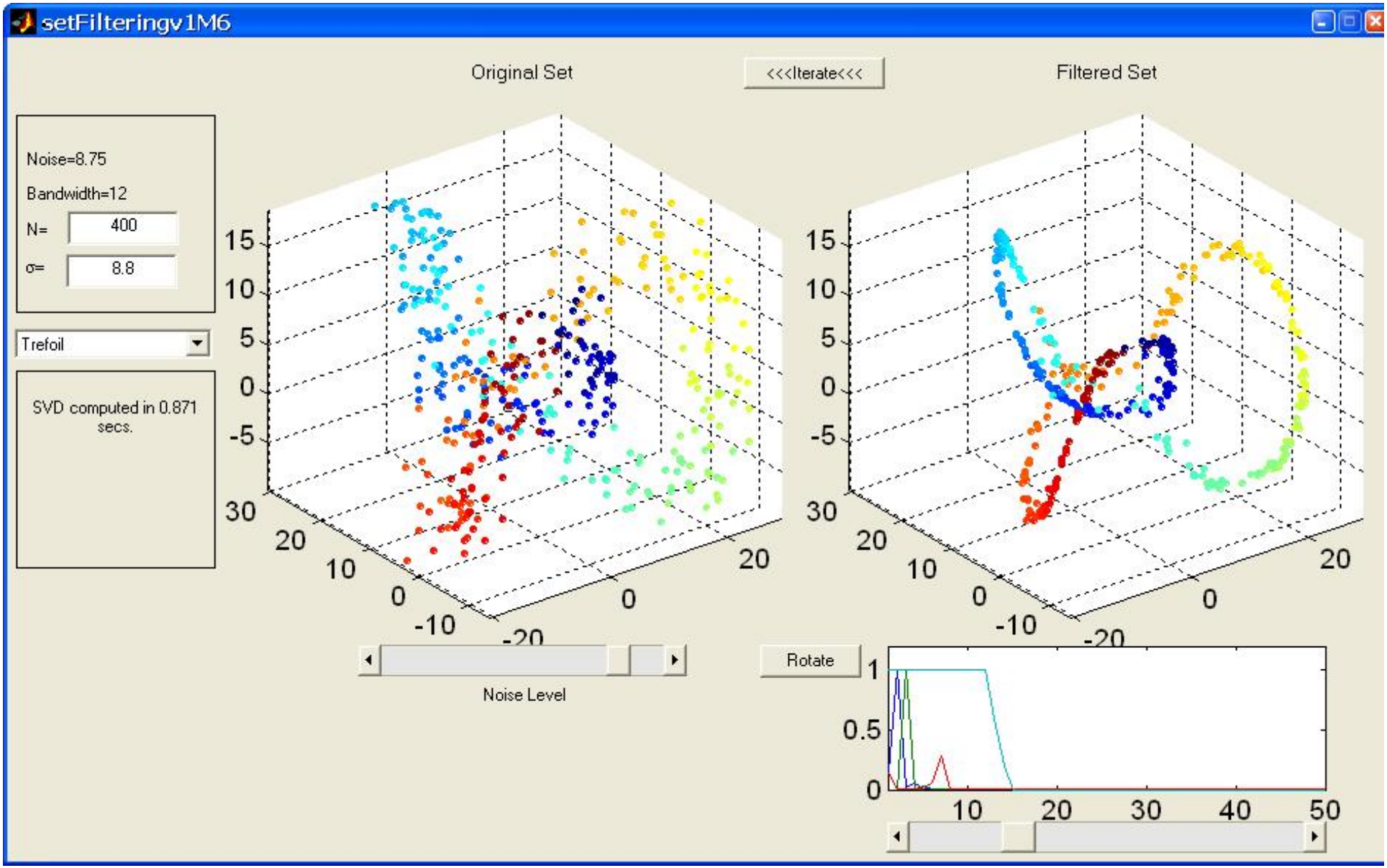
•*Data exploration, and perceptualization is enabled by the diffusion map since it converts complex inference chains to ordinary physical distance in the perceptual displays,* ***to provide situational awareness.***

•*The diffusion geometry which is induced by the various chains of inference enable a* ***multiscale hierarchical organization*** *governed by the scaling of the length of chains.*

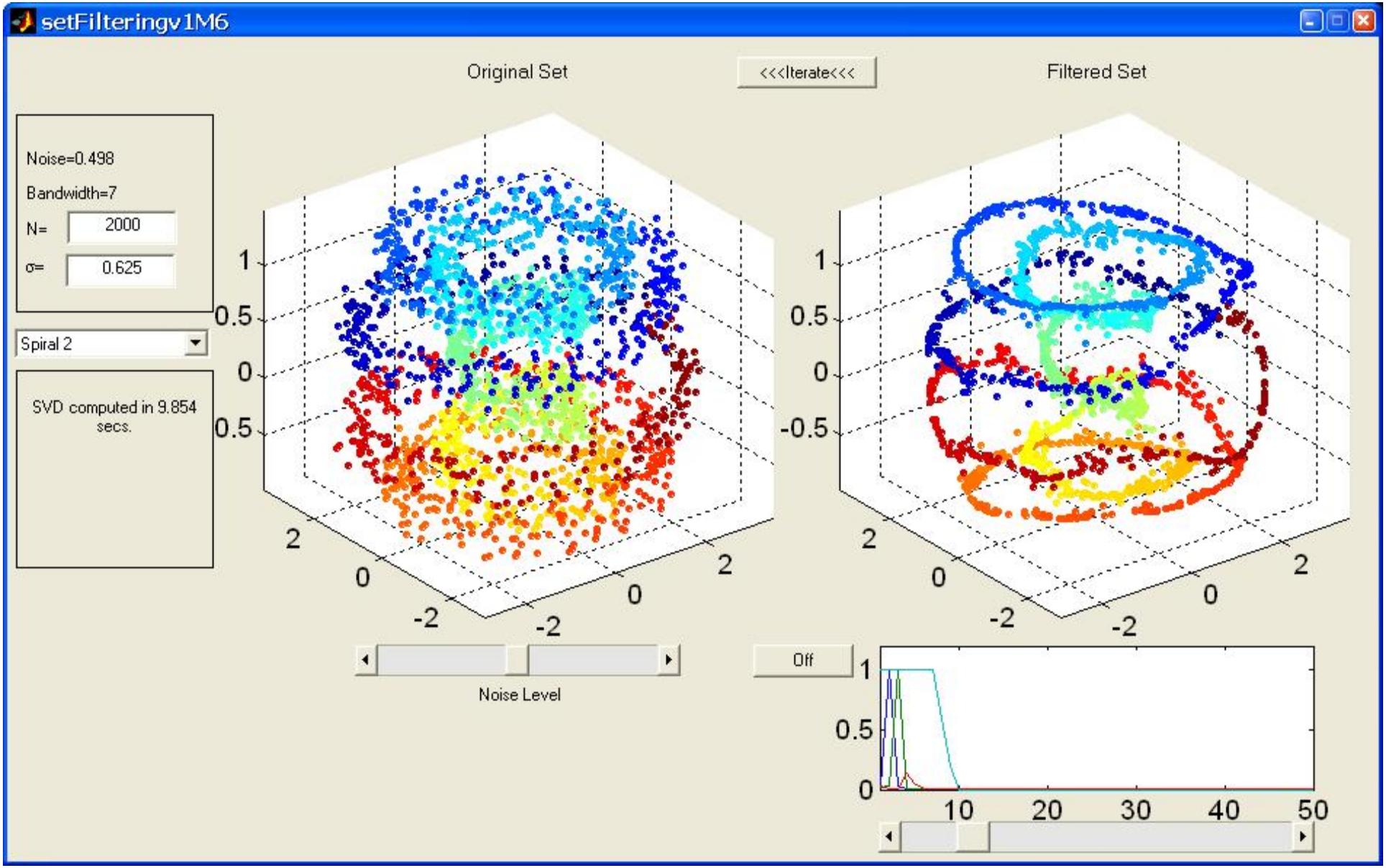
•*Diffusion from reference points provides efficient linkages to related points propagating searches into the data while* ***ranking for relevance***

Intrinsic data cloud filtering by diffusion geometry.

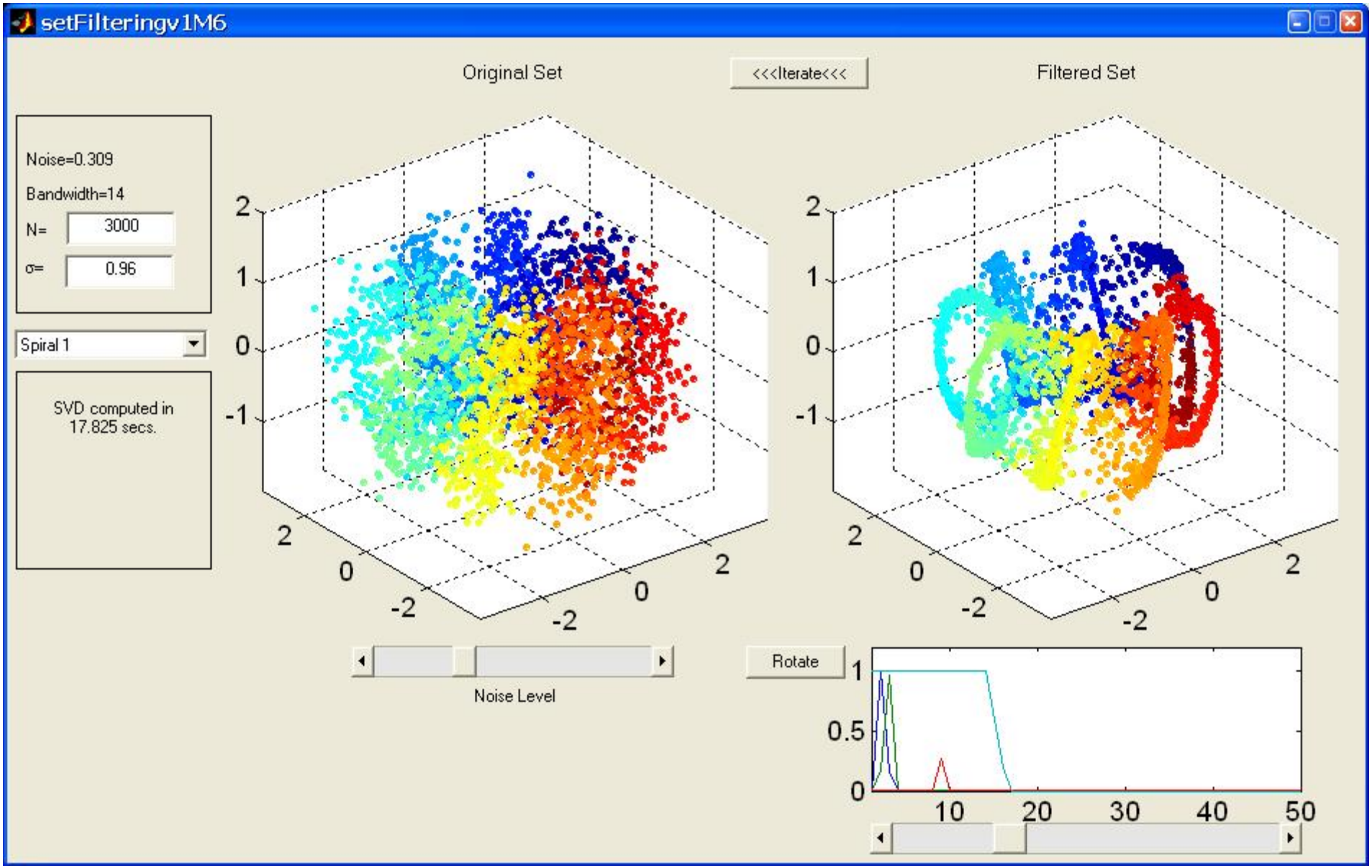
This method extends the algorithm in which the curve is parametrized by arc length and the filtered using Fourier modes. In our case the eigenfunctions are discretized FT approximations



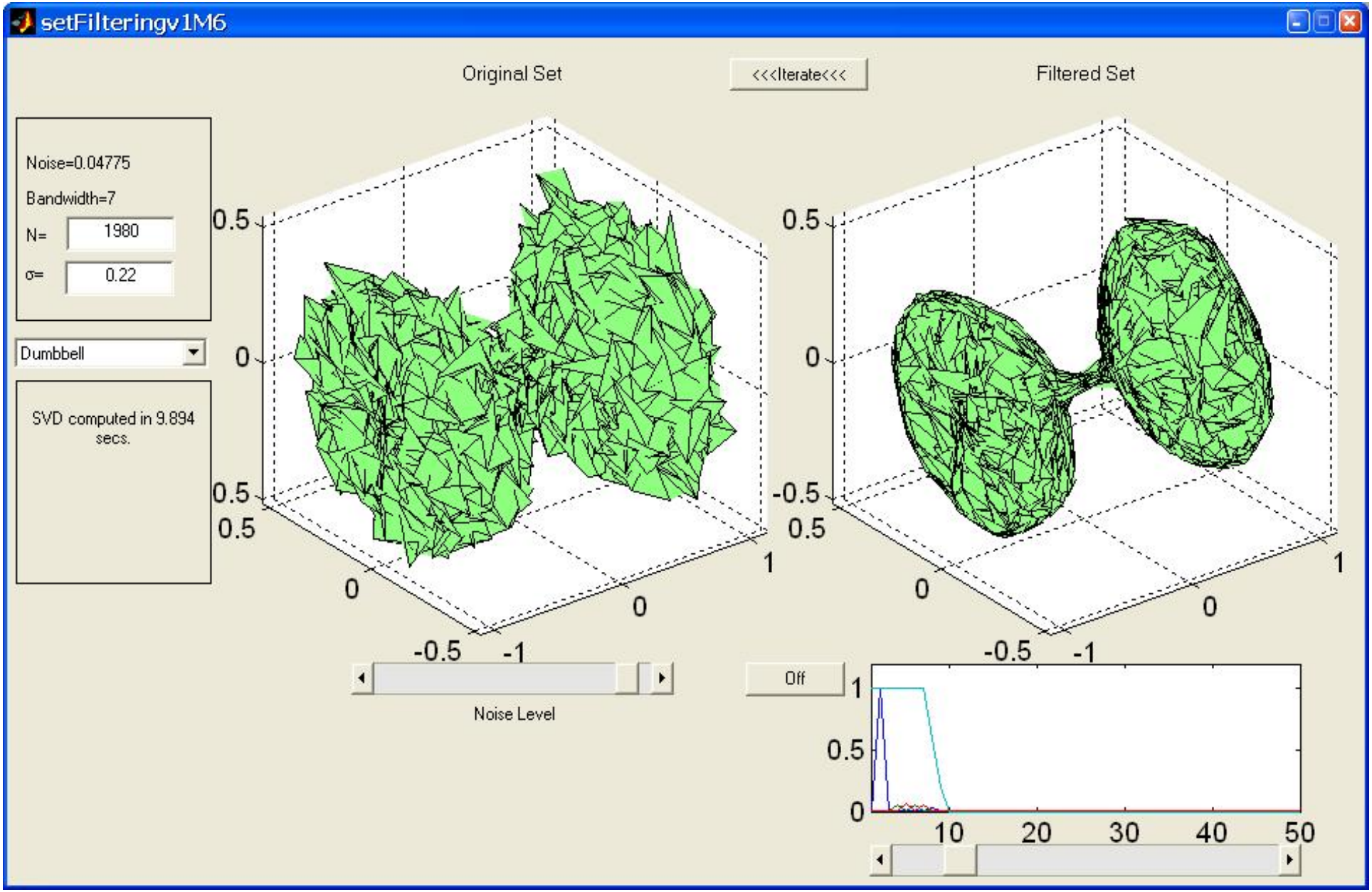
The data is given as a random cloud , the filter organizes the data. *The colors are not part of the data*



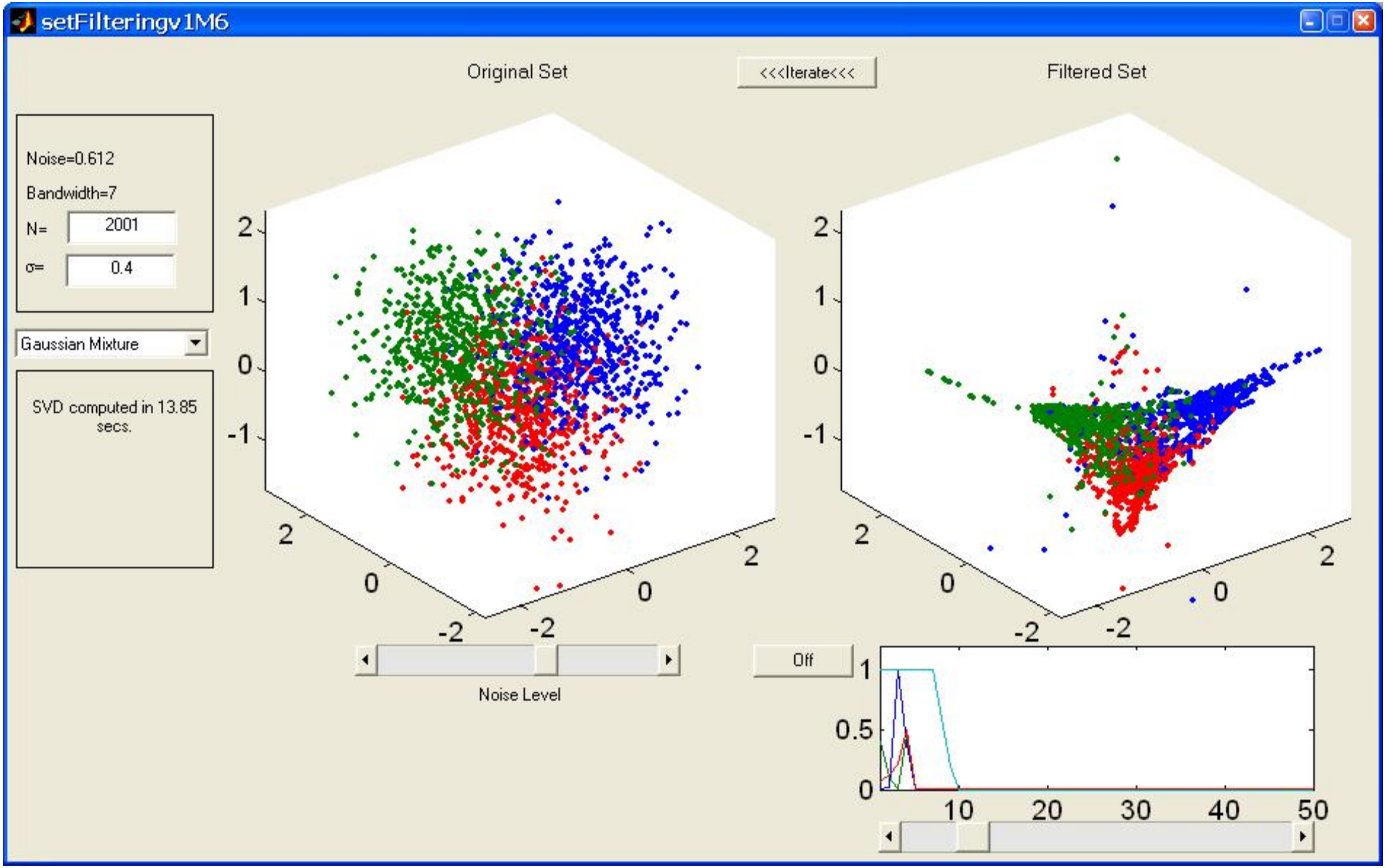
A noisy disorganized spiral , filtered at band 14 .



Point cloud around a dumbell surface



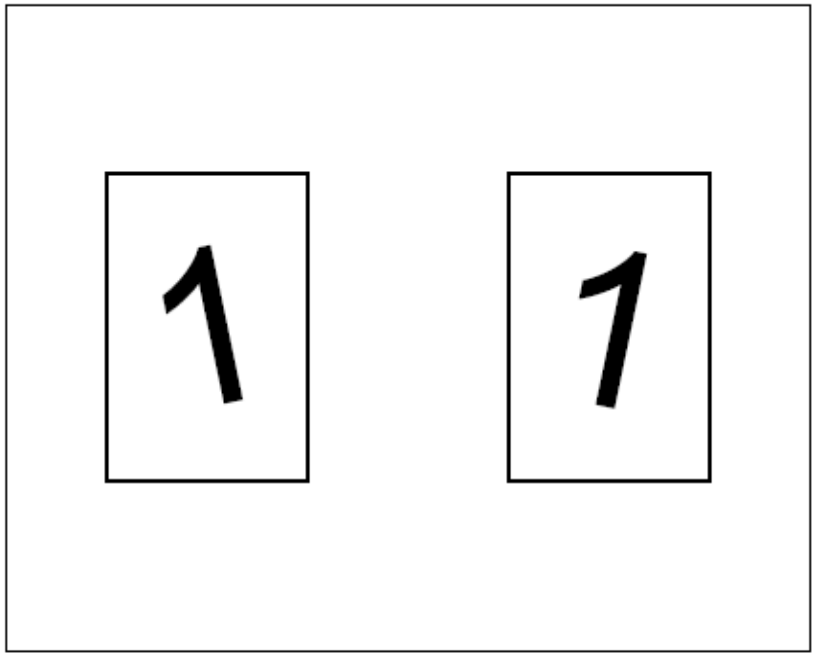
Gaussian cloud filtering by diffusion



Problems in High Dimensions

Many algorithms are based on distances $\|\boldsymbol{x}_i - \boldsymbol{x}_j\|$

In high dimensions, large distances are (almost) meaningless.



In many applications, high dimensional data has an intrinsic low dimensionality.

dimensional reduction: How can we embed data in a low dimensional space (and find the intrinsic dimensionality) ?

Diffusion as a search mechanism. Starting with a few labeled points in two classes , the points are identified by the “preponderance of evidence”. (Szummer ,Slonim, Tishby…)

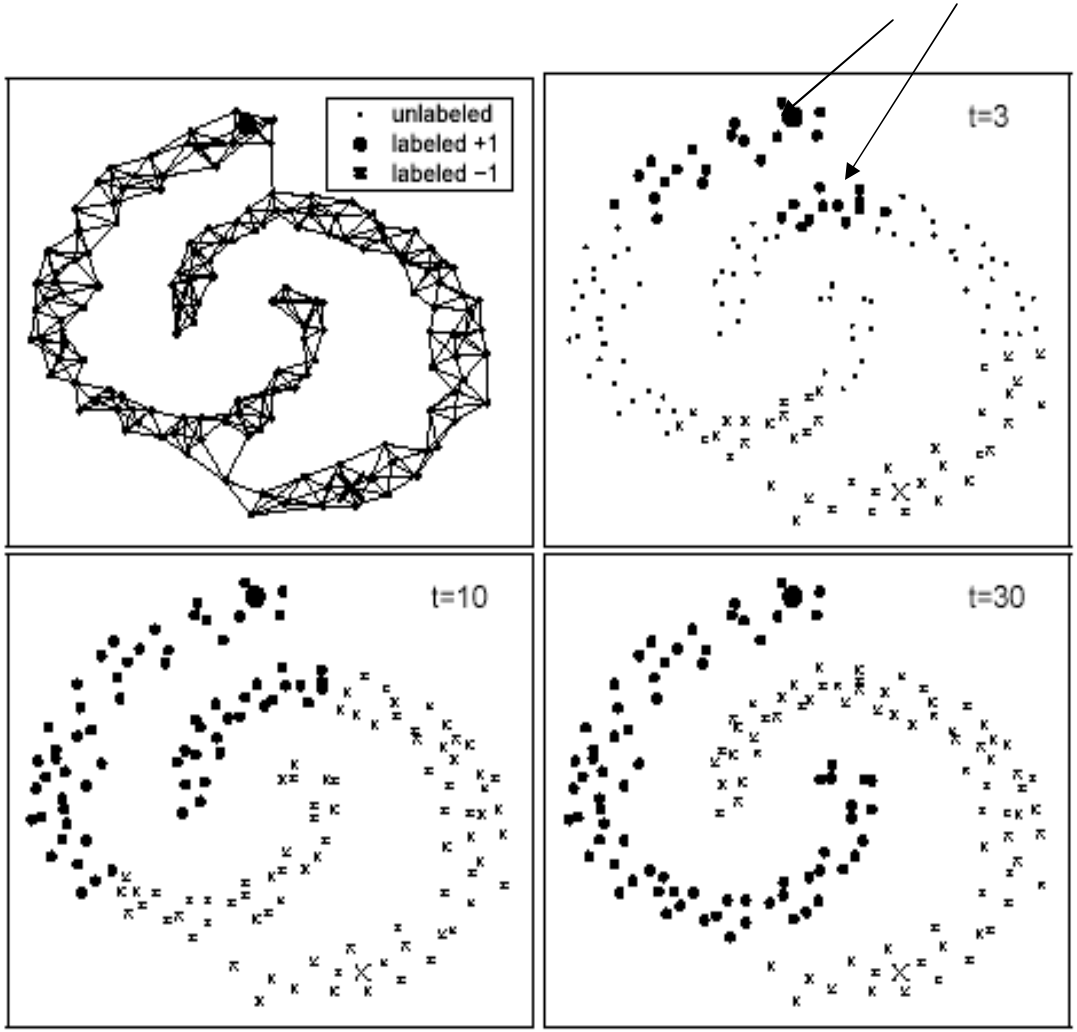
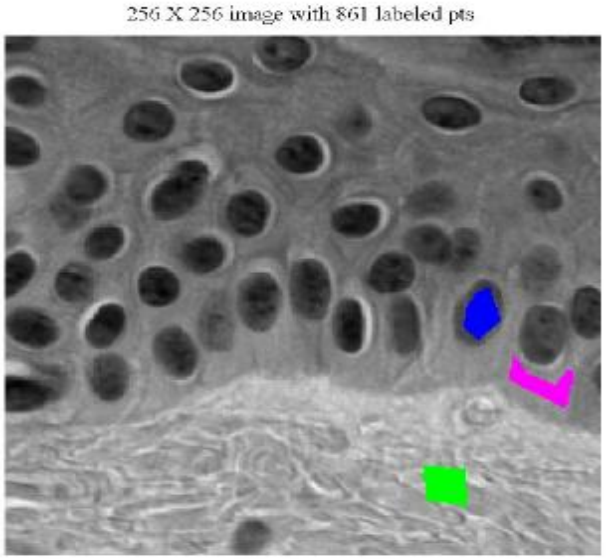
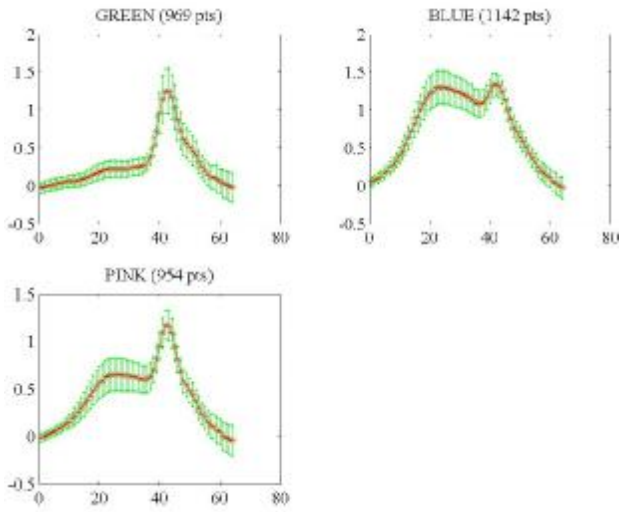
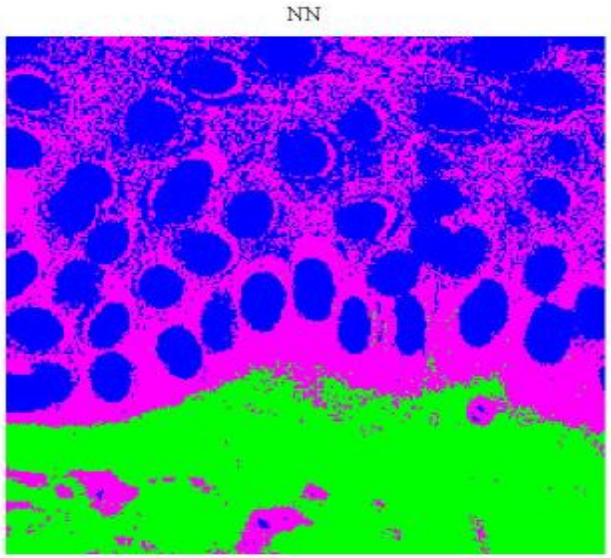


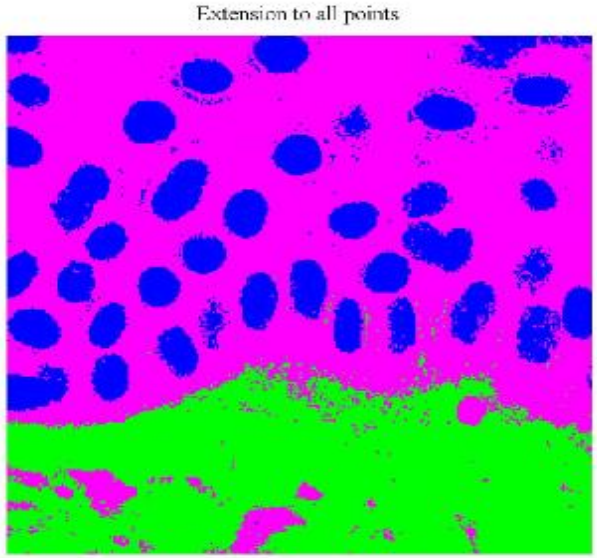
Figure 1: Top left: local connectivity for K=5 neighbors. Below are classifications using Markov random walks for t=3, 10, and 30 (top to bottom, left to right), estimated with average margin. There are two labeled points (large cross, circle) and 148 unlabeled points, classified (small crosses, circles) or unclassified (small dots).

Conventional nearest neighbor search , compared with a diffusion search. The data is a pathology slide ,each pixel is a digital document (spectrum below for each class)

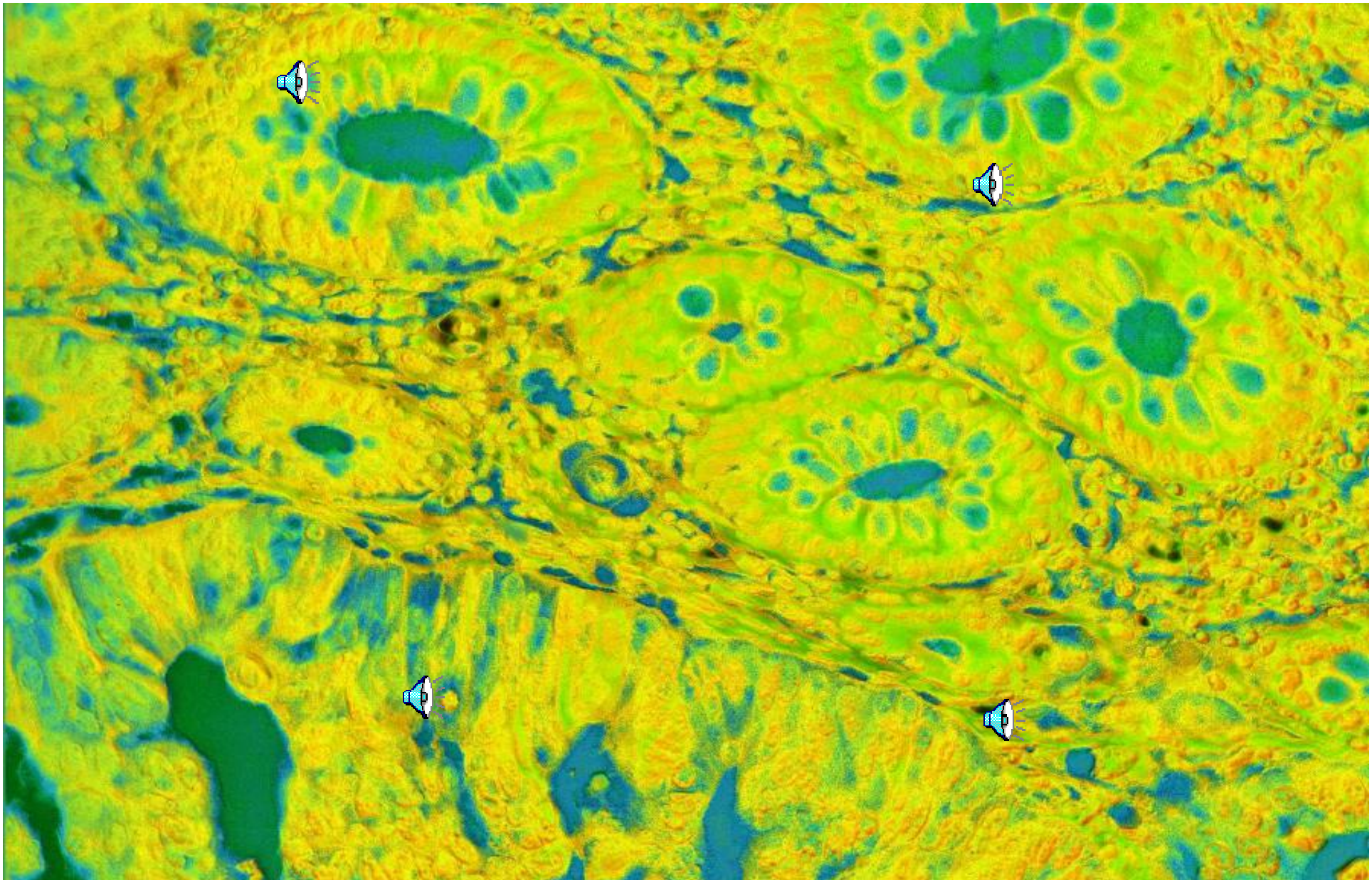


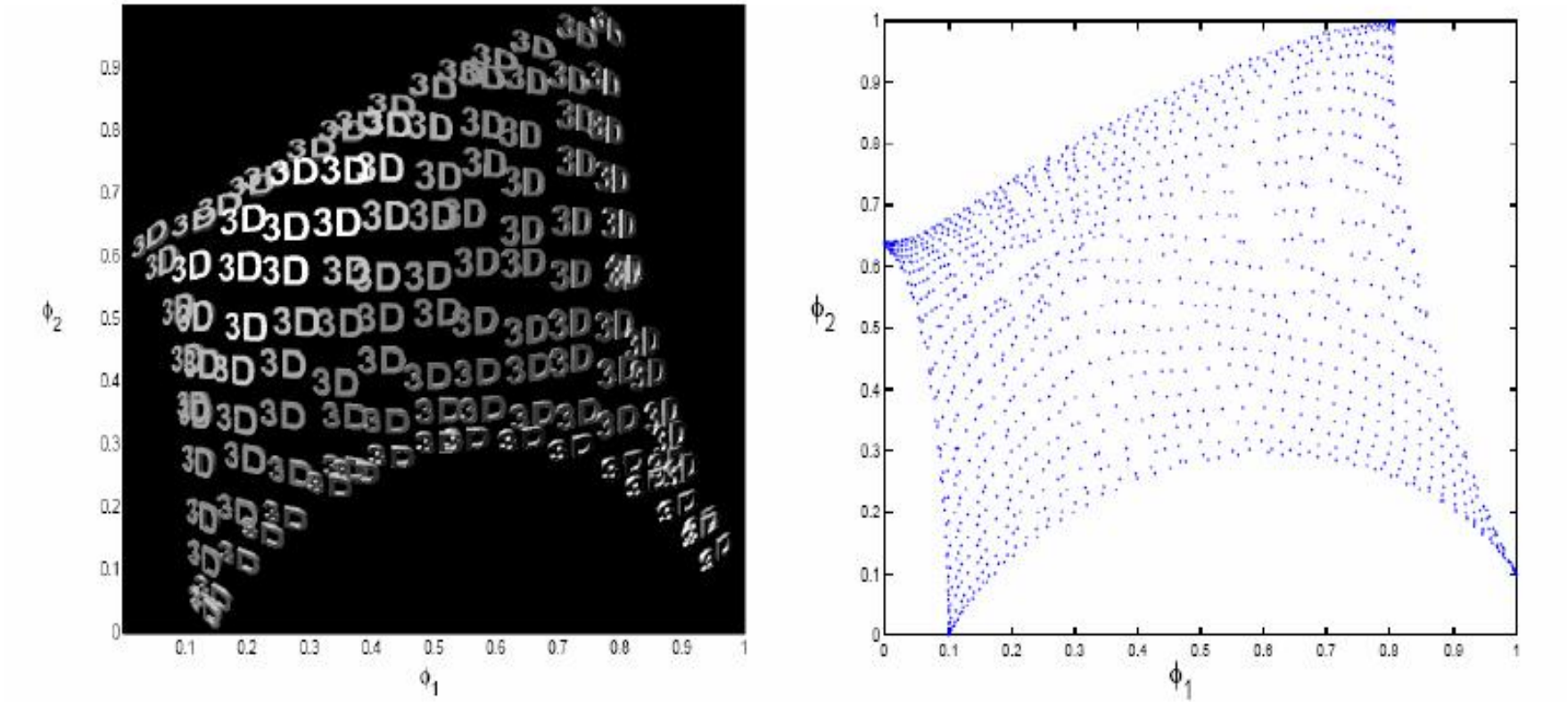






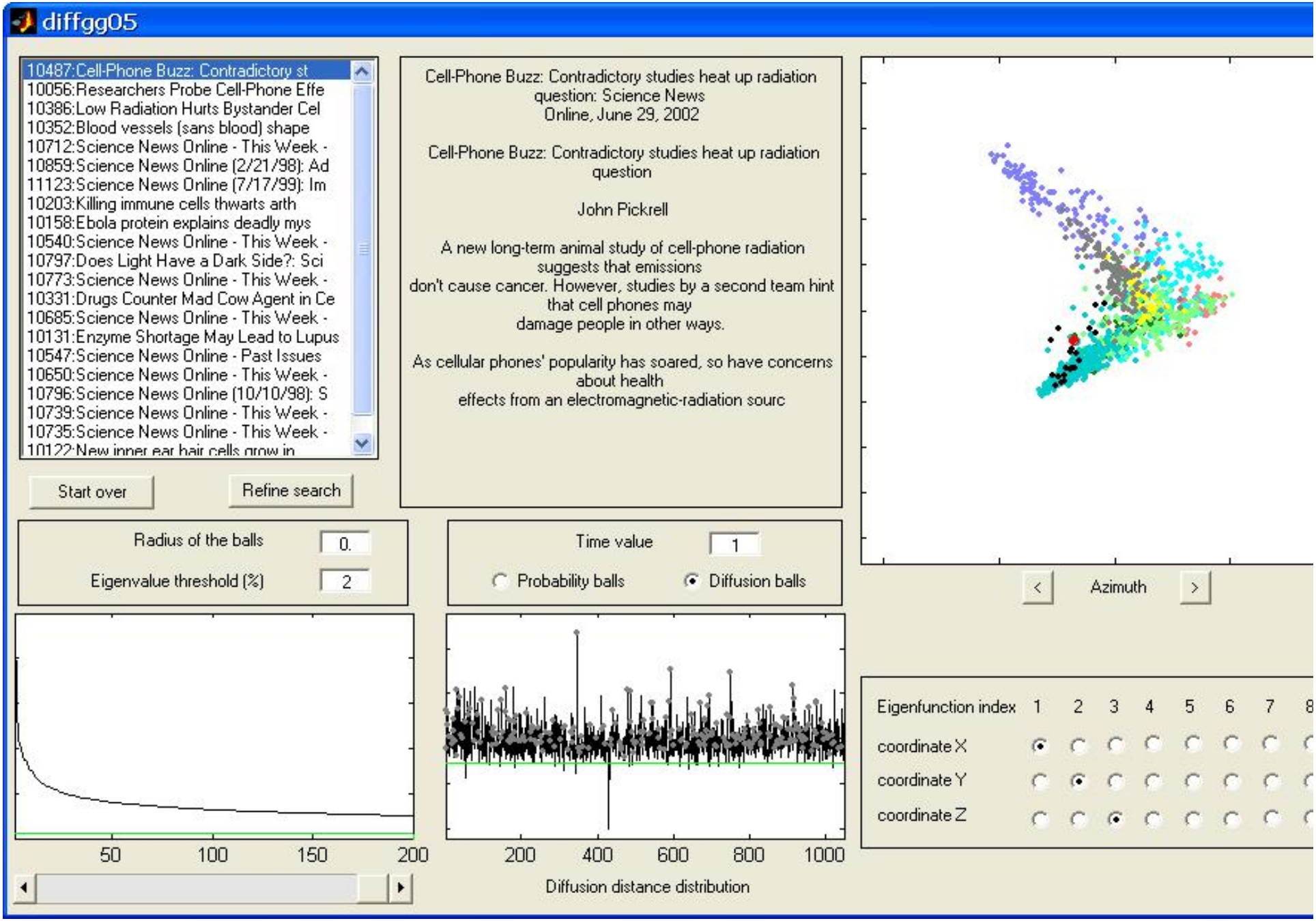
A pathology slide in which three tissue type are displayed as an RGB mixture ,where each color represents the probability of being of a given type .





The First two eigenfunctions organize the small images which were provided in random order

Organization of documents using diffusion geometry



A simple model for the generation of point clouds * is provided through the Langevin and Fokker-Planck equations :

Consider a stochastic differential equation in $n \gg 1$ dimensions

$$\dot{\boldsymbol{x}} = -\nabla U(\boldsymbol{x}) + \dot{\boldsymbol{w}}$$

where

$\boldsymbol{x}(t) \in \mathbb{R}^n$ - configuration at time t

$U(\boldsymbol{x})$: potential

$\boldsymbol{w}(t)$: n-dimensional Brownian motion

Examples: macromolecules and proteins in water, interacting particle systems, low dimensional projections of deterministic systems

*----------

Think of propagation of fire in the presence of wind. Or water draining down a landscape

Let $p(\boldsymbol{x}, t|\boldsymbol{y}, t_0)$ denote the forward transition probability density ($t > t_0$).

It satisfies the Fokker-Planck (FP) equation

$$\frac{\partial p}{\partial t} = \nabla \cdot [\nabla p + p \nabla U]$$

The steady state follows the Boltzmann distribution

$$p(\boldsymbol{x}) = e^{-U(\boldsymbol{x})}/Z, \qquad Z - \text{normalization factor}$$

Approach to steady state governed by the eigenfunctions with the smallest eigenvalues of the FP operator.

When $n \gg 1$ the FP equation cannot be solved numerically or analytically. However, simulations of the SDE are "easily performed"

Weighted Diffusion Maps

If instead we define a new weighted kernel

$$\tilde{k}_\varepsilon(\boldsymbol{x}, \boldsymbol{y}) = \frac{k_\varepsilon(\boldsymbol{x}, \boldsymbol{y})}{\sqrt{p_\varepsilon(\boldsymbol{x}) p_\varepsilon(\boldsymbol{y})}}$$

and consider the normalized graph Laplacian (diffusion map) based on this kernel, then as $\varepsilon \to 0$, the corresponding evolution equation is

$$\frac{\partial p}{\partial t} = \tilde{\mathcal{H}}_f p = \Delta p - \nabla p \cdot \nabla U$$

This equation is the same as the backward FP equation of the SDE

$$\dot{\boldsymbol{x}} = -\nabla U + \dot{\boldsymbol{w}}$$

Time Evolution

Approach to steady state governed by the eigenfunctions with the smallest eigenvalues of the FP operator.

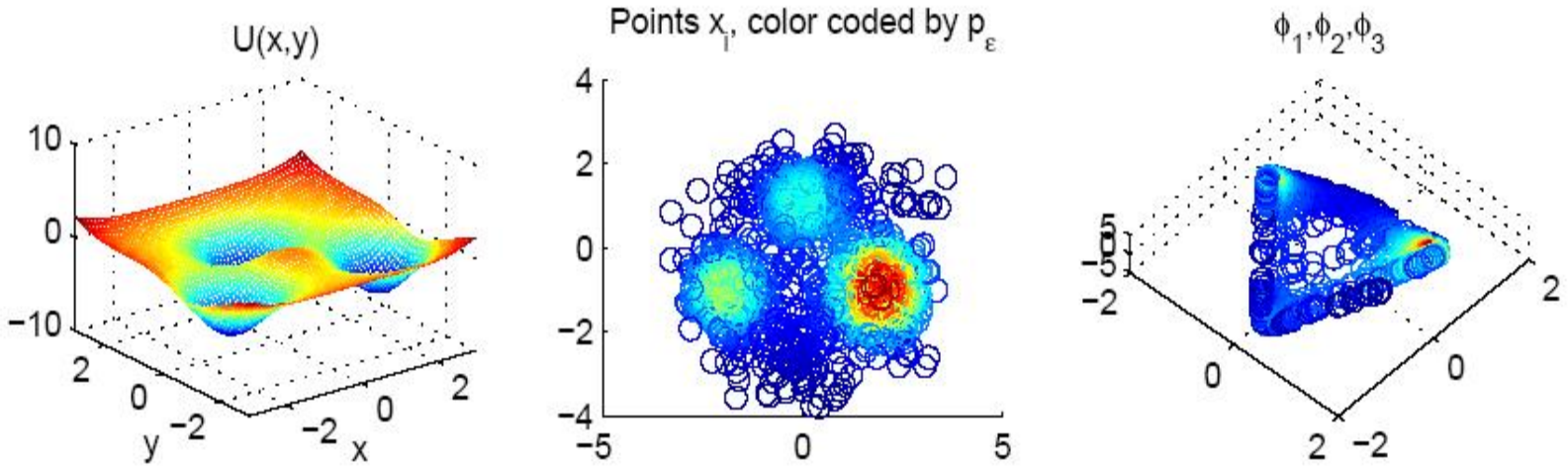
$$p(\boldsymbol{x}, t) = \sum_j a_j e^{-\lambda_j t} \psi_j(\boldsymbol{x})$$

When $n \gg 1$ the FP equation cannot be solved numerically or analytically. However, simulations of the SDE are "easily performed"

Given a lot of simulated data $\{\boldsymbol{x}_i\}_{i=1}^N$,$N \gg 1$, possibly from many different trajectories, can we approximate the first few eigenfunctions $\psi_j(\boldsymbol{x})$?

3-well potential

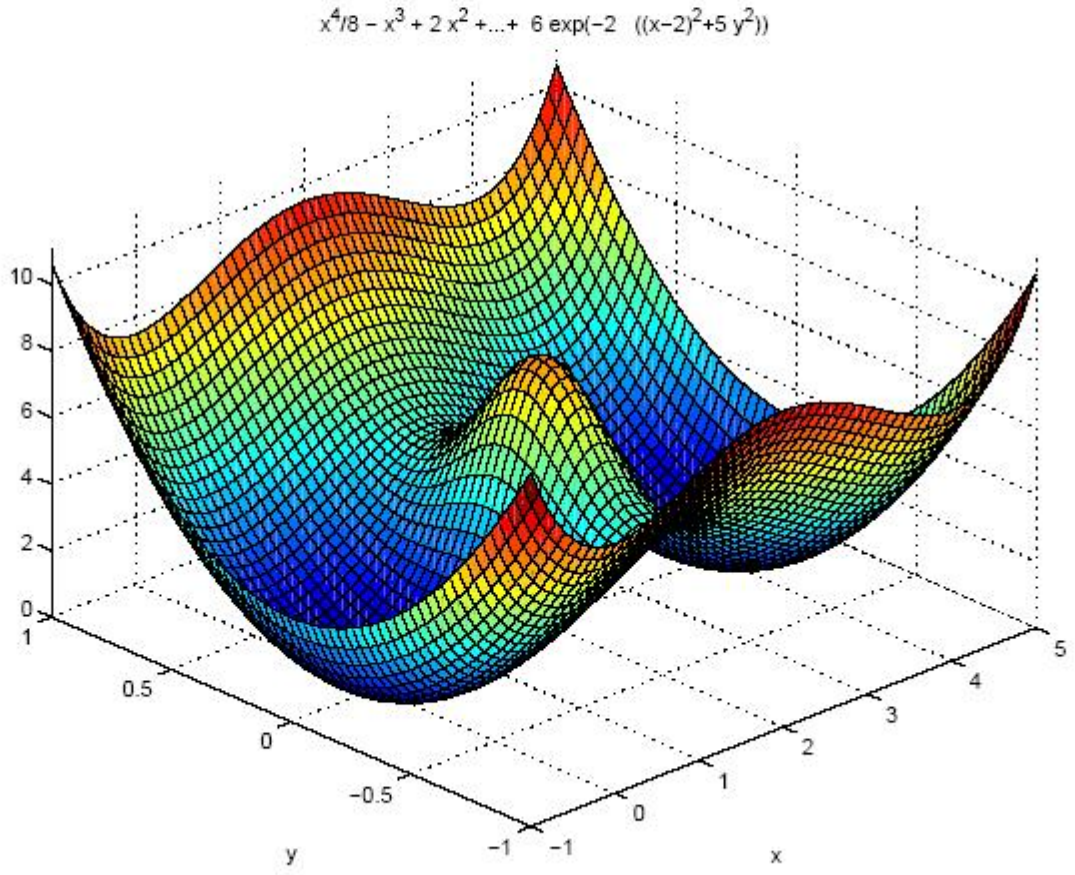
Case I: Well defined potential wells

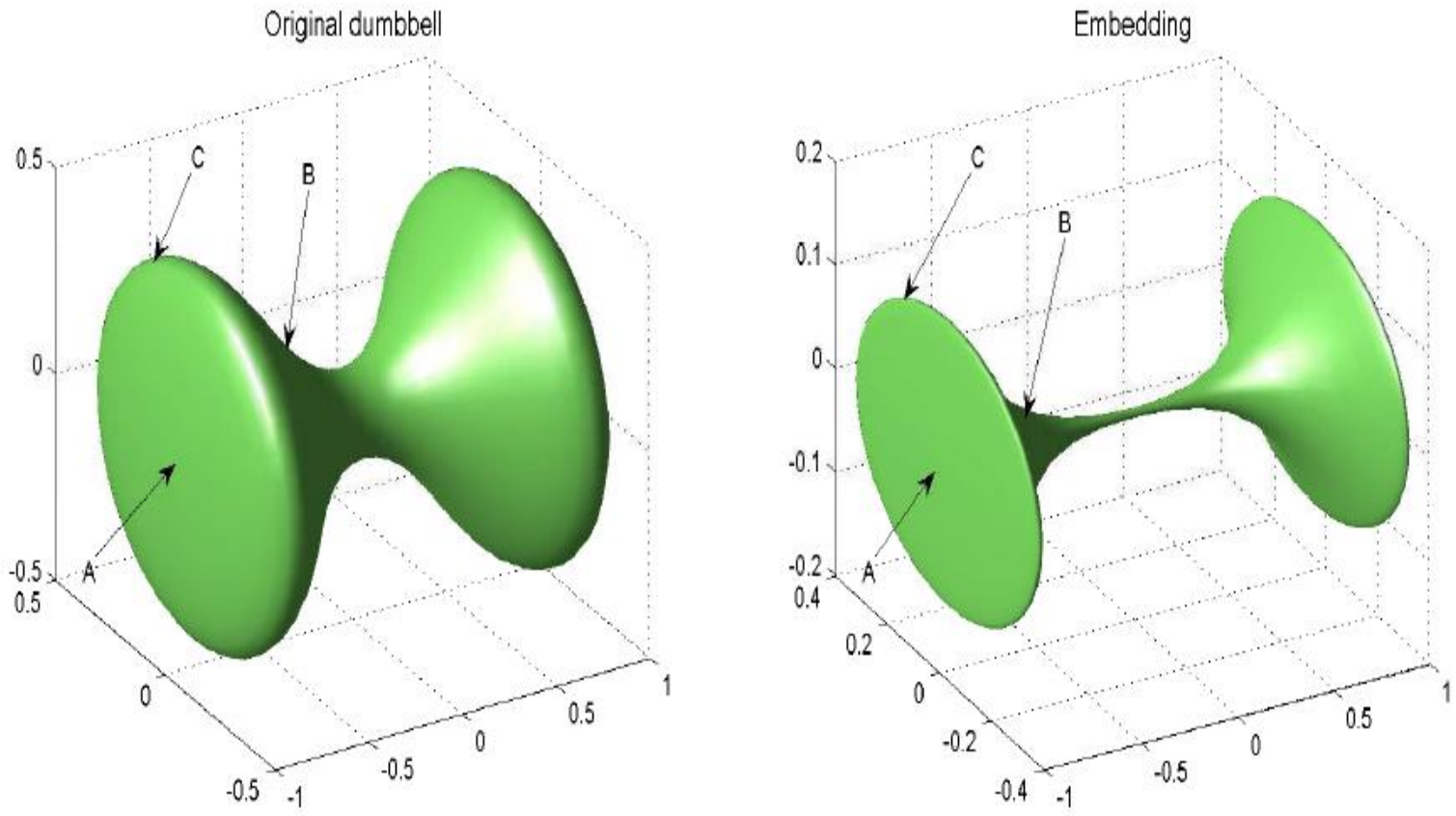


2-D diffusion map clearly shows the 3 wells and the most common paths between them.

Double Well Potential with 2 connecting paths

$$U(x,y) = \left(\frac{x^4}{8} - x^3 + 2x^2 + \frac{y^2}{2\tau_2} + 6\exp(-2(x-2)^2 - 10y^2) \right)$$





The natural diffusion on the surface of the dumbell 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.

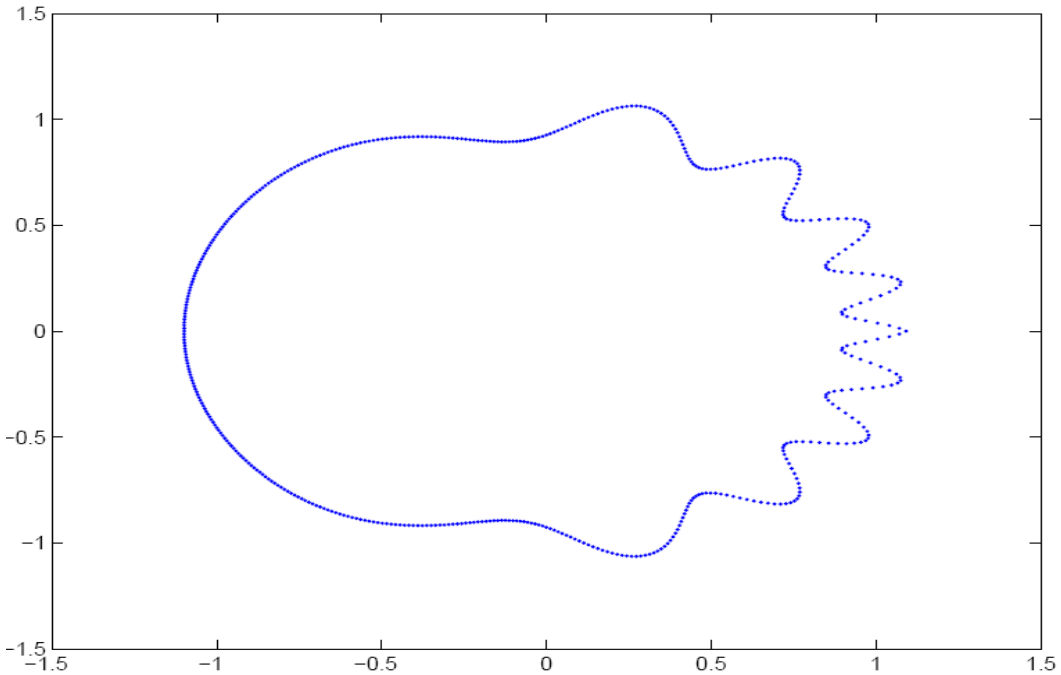
Figure 2.9: Left: Set of images randomly permuted. This is the input of the algorithm. Right: output of the algorithm, the sequence is reordered with respect to the angle of rotation of the head (the sequence is to be read from left to right, and top down).

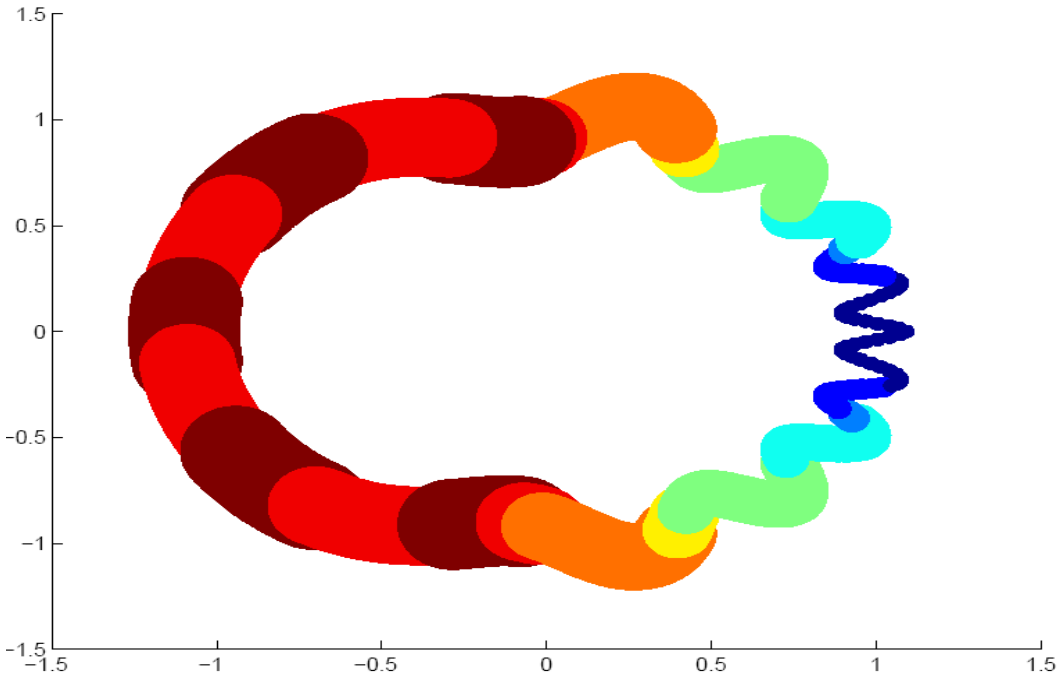
Extension of Empirical functions off the data set .

An important point of this multiscale analysis involves the relation of the spectral theory on the set to the localization on and off the set of the corresponding eigenfunctions .

In the case of a smooth compact submanifold of Euclidean space it can be shown that any band limited function of band B can be expanded to exponential accuracy in terms of eigenfunctions of the Laplace operator with eigenvalues not exceeding B^2

Conversely every eigenfunction of the laplace operator satisfying this condition extends as a band limited function with band C'B (both of these statements can be proved by observing that for eigenfunctions of a Laplace operator we can estimate the size of a derivatives of order 2m as a power m of the eigenvalue, implying that eigenfunctions on the manifold are well approximated by restrictions of band limited functions of corresponding band.





Multiscale extension of expansions on the data set . The bottom image shows the distance to which low degree expansions in Laplace eigenfunctions can be extended reliably.

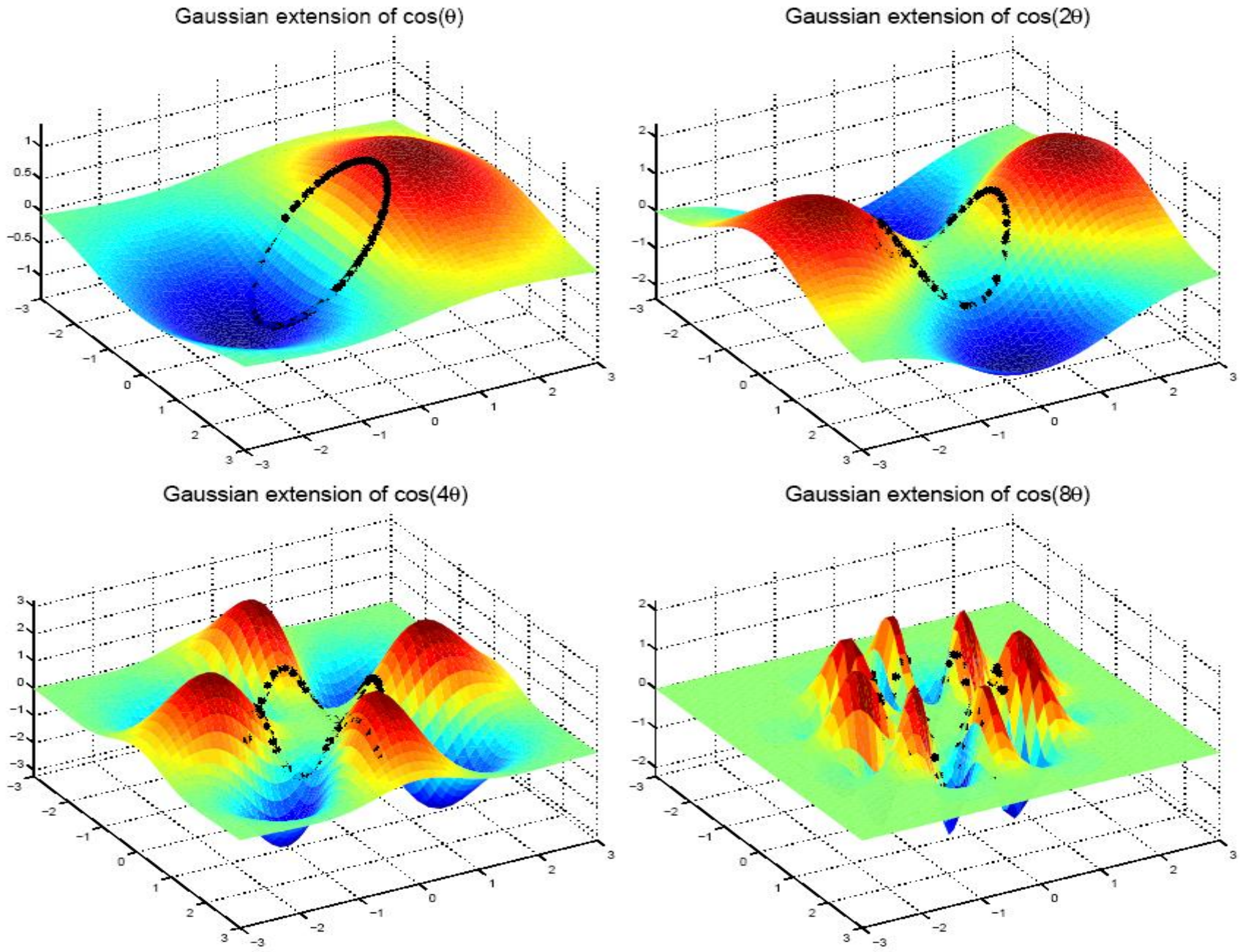


FIGURE 6. Extension of the functions $f_j(\theta) = \cos(2\pi j\theta)$ for $j = 1, 2, 3$ and 4, from the unit circle to the plane.