Conference in Honor of Aline Bonami
Orleans, June 2014

Harmonic Analysis and functional duality , as a tool for organization of information, and learning.

R. Coifman
Department of Mathematics,
program of Applied Mathematics
Yale University
Joint work with J. Ankerman, M. Gavish, A. Haddad, W. Leeb

• Harmonic Analysis has over the last 60 years focused on the relationship between geometry , and appropriate representations, as a tool to understand and prove estimates on operators . In particular kernels of operators restricted to subsets of Euclidean space have played a fundamental role in understanding the geometry and combinatorics of the set.

•We claim that these methodologies open the door to organization of matrices viewed as either databases, or as linear transformations.

The challenge is to organize a database or a matrix without any a priori knowledge of its internal model, in particular can we find data anomalies, fill in missing entries build classifiers and in general build data agnostic, analytic mathematics for processing any kind data.

•Agnostic data geometerization, enables automation of data organization and fusion +analytical intelligence.
Like a good memory organization, we would have the first step to ab initio learning, learning in which we have a feedback mechanism to reorganize the data according to the inferences we wish to achieve.

- The main analytical challenge is to simultaneously build a graph of columns and a graph of rows so that the matrix entries are as smooth (or predictable)as possible, relative to the tensor product of these geometries. This smoothness is measured in terms of an appropriate tensor Besov norm or entropy .

- The next challenge is to enable simple reorganization to achieve regression or machine learning, or fast numerical analysis.

We illustrate the outcome of such organization on the MMPI (Minnesota Multiphasic Psychological Inventory) questionnaire .

The underlying analytical methods enables filtering out anomalous responses , and provides detailed quantitative assessments of consistency of responses .

The analysis-synthesis tools, that enable the geometric construction, are useful to provide a metric to assess success in organizing the data base.

We extend ideas of Harmonic Analysis and approximation theory to the study of general matrices , whether the goal is organization of a data base to extract knowledge, or to build a representation relative to which a matrix is efficiently described.
We illustrate the outcome of such organization on the MMPI (Minnesota Multiphasic Psychological Inventory) questionnaire .
The Tensor Haar Bases enable filtering out anomalous responses , and provide detailed "analysis" (pun intended) .

Stromberg's observations about the efficiency of approximation of functions of bounded mixed variation in the tensor Haar basis is particularly useful in the statistical data analysis context of analysing a data base

Start by considerimg the problem of unraveling the geometric structure in a matrix. We view the columns or the rows as collections of points in high dimension whose geometry we need to define.

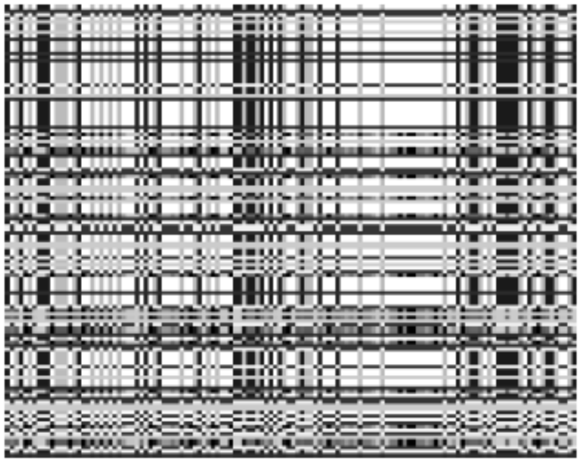
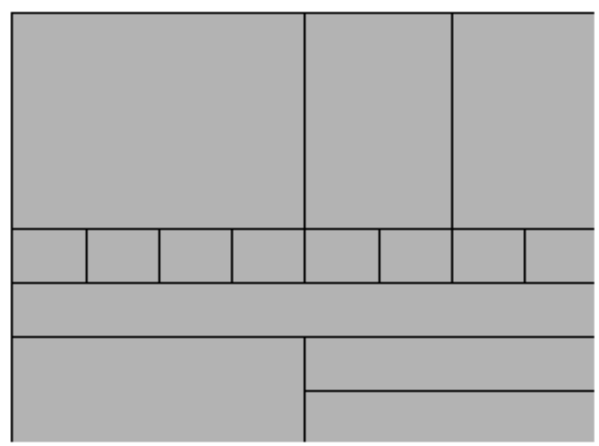


Figure 4.2: A permutation of the matrix A.

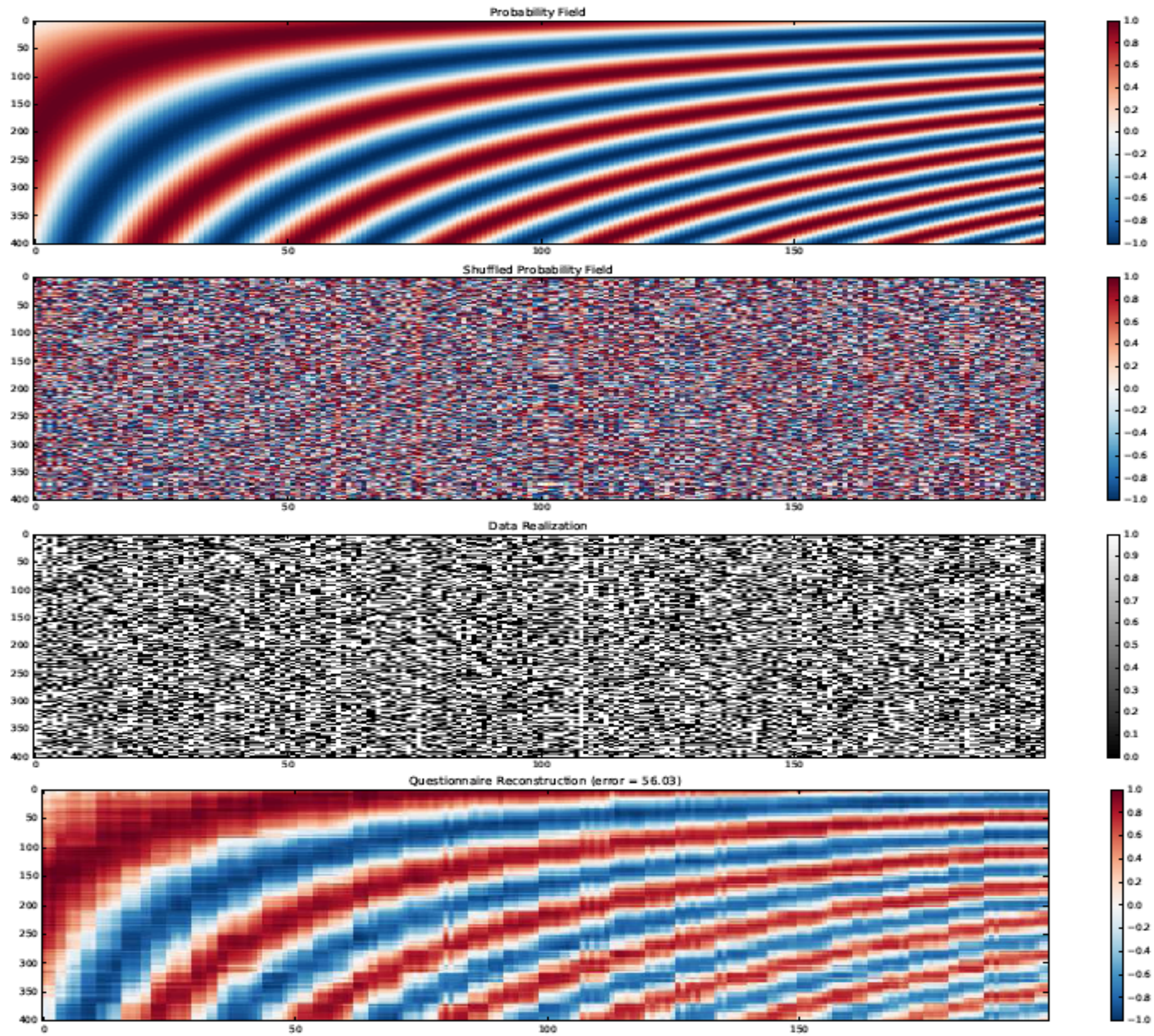
The matrix on the left is a permutation in rows and columns of the matrix below it .

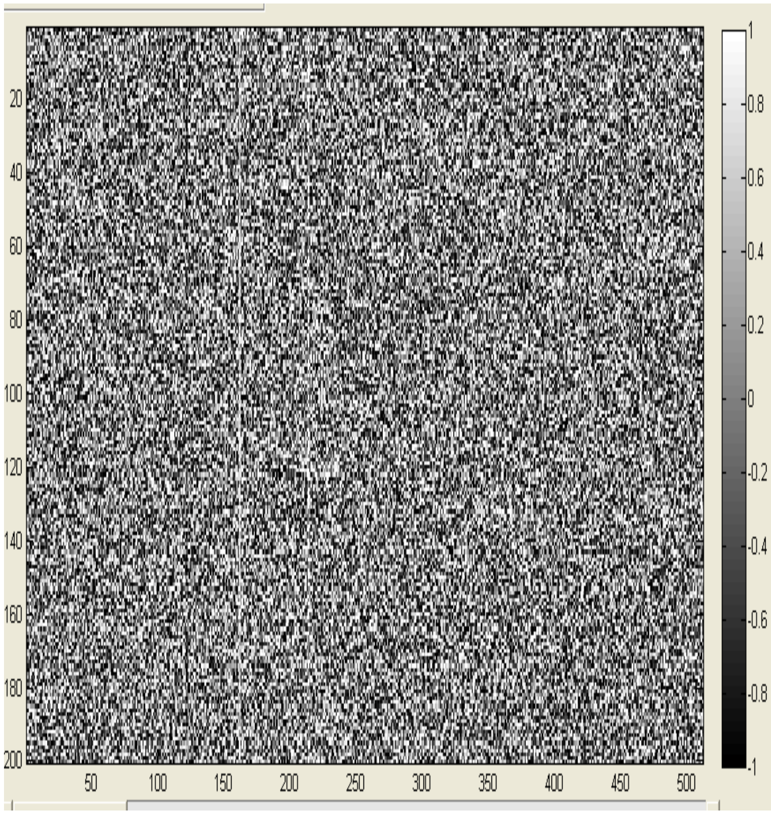
The challenge is to unravel the various simple submatrices .



More generally assume that the function represents a probability field which has be garbled by permuting rows and columns. At each pixel we toss a coin with corresponding probability .

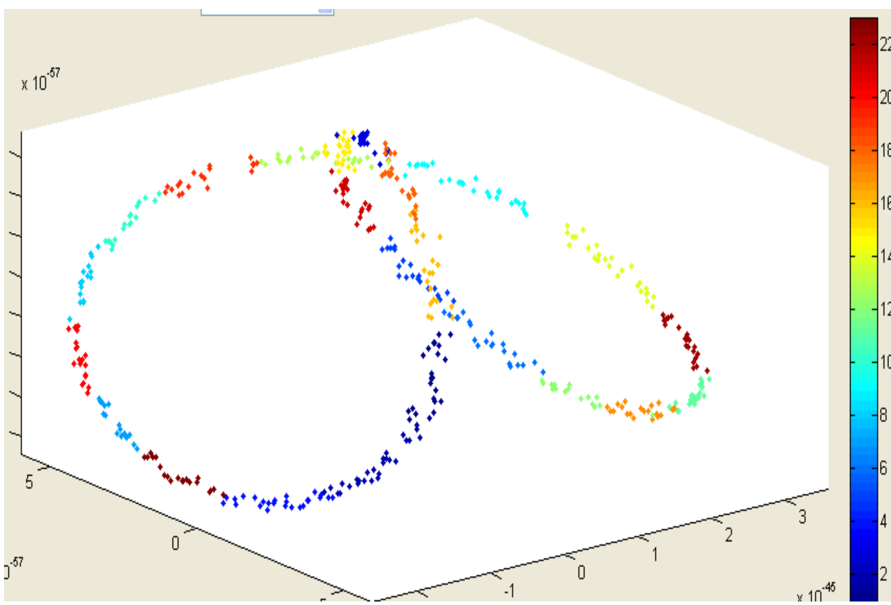
The Challenge is to recover the underlying field with some accuracy control.

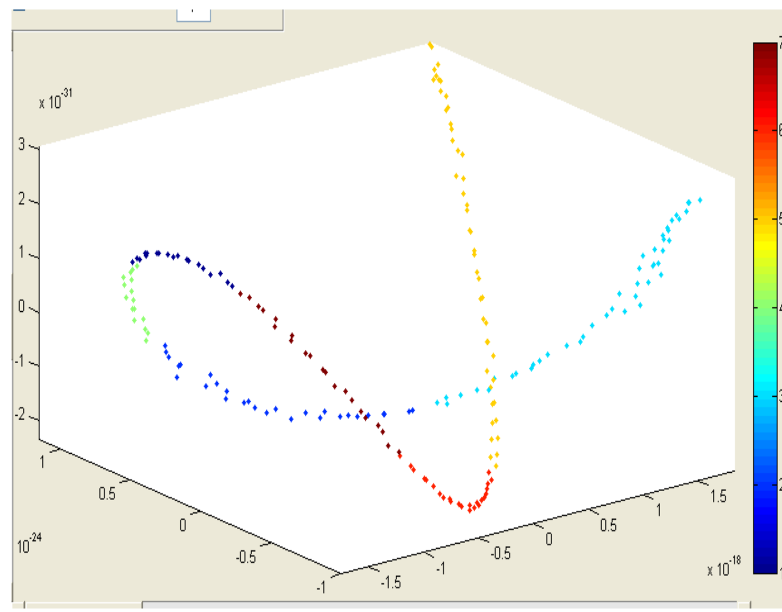




A permutation of the rows and columns of the matrix sin(kx). On the left we recover the one dimensional geometry of x (which is oversampled), while on the right we recover the one dimensional geometry of k .

More generally we can build a dual geometry of eigenvectors of Laplace Beltrami operators on manifolds





The simplest joint organization is achieved as follows

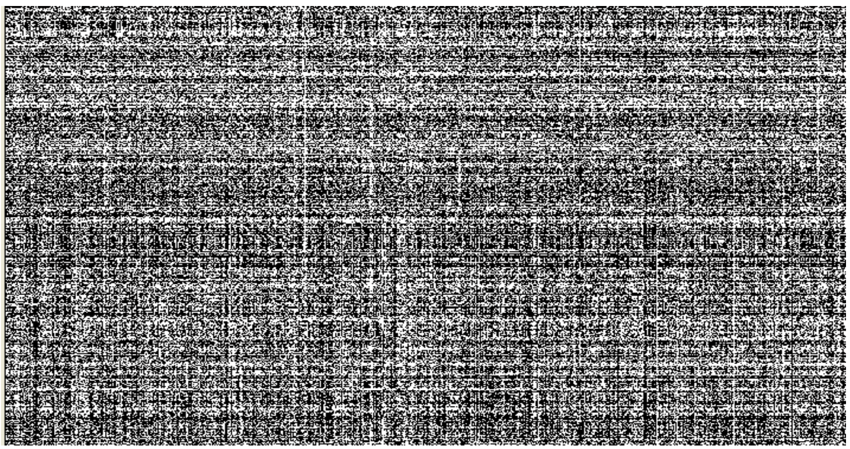
Assuming an initial hierarchical organization of the columns of the database (see later) into contextual folders (for example groups of responders which are similar at different “scales”) use these folders to assign new response coordinates to each row (question), for example an average response of the demographic group.

Use the augmented response coordinates to organize responses into a conceptual hierarchy of folders of rows which are similar across the population of columns.

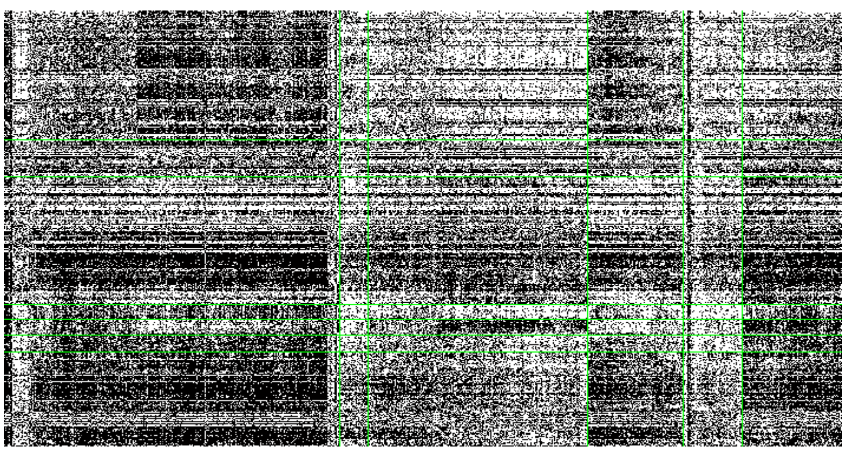
We then use the conceptual folders to augment the response of the columns and to reorganize them into a more precise contextual hierarchy .

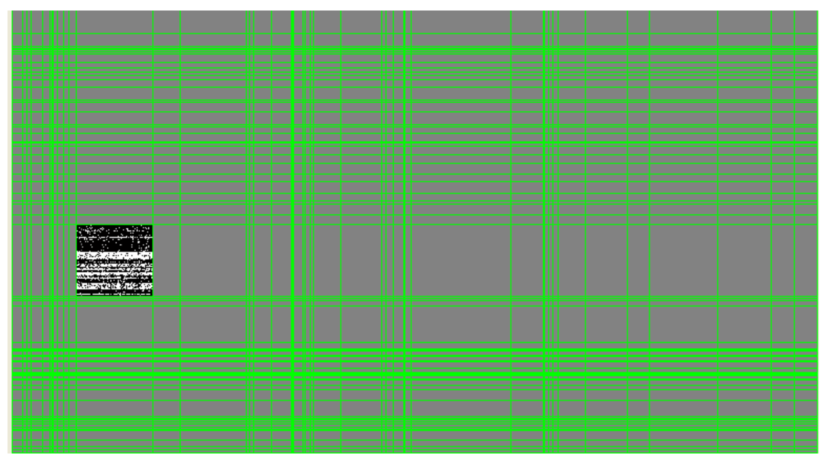
This process is iterated as long as an “entropy “ of the database is being reduced .

The challenge is to organize a data base by organizing both rows and columns simultaneously , if the columns are observations and the rows are features or responses. We organize observations “contextually” and responses “conceptually “ each organization informs the other iteratively.



A disorganized questionnaire ,on the left, the columns represent people , the row are binary questions. Mutual multiscale bi learning , organizes the data, bottom left , The questionnaire is split on a two scale grid below. Showing in the highlighted rectangle , the consistency of responses of a demographic group (context) to a group of questions (concept)





Consider the example of a database of documents , in which the coordinates of each document , are the frequency of occurrence of individual words in a lexicon. Usually the documents are assumed to be related if their vocabulary distributions are “close” to each other.
The problem is that we should be able to interchange words having similar meaning and similarity of meaning should be part of the document comparison .
By duality if we wish to compare two words by conceptual similarity we should look at similarity of frequency of occurrence in documents, here again we should be able to interchange documents if their topical difference is small.
There are at least three challenges which we claim can be resolved through Harmonic Analysis ;

1. Define good document content flexible-distances , and simultaneously good conceptual vocabulary distances.

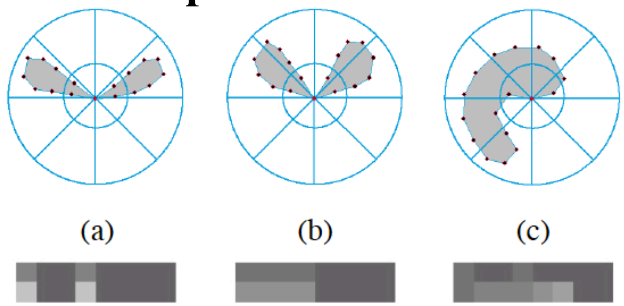
2. Develop a method which is purely data driven and data agnostic ,

3.The complexity of calculations should scale linearly with data size.

We start by discussing metrics

YOSSI RUBNER, CARLO TOMASI AND LEONIDAS J. GUIBAS

P. Indyk and N. Thaper. **Fast image retrieval via embeddings.**



(a) (b) (c)

(d) (e) (f)

Distance	$d(a,b)$	$d(b,c)$
L_1	1.0	0.875
L_2	0.3953	0.3644
χ^2	0.6667	0.6625
EMD	0.5	1.5625

(g)

From Sameer Shirdhonkar and David W. Jacobs

Discrete EMD for histograms	Continuous EMD for distributions
Histograms $f(i;1), f(i,2)$ $\sum_i f(i;1) = \sum_i f(i;2) = 1$ Difference $f(i) := f(i;1) - f(i;2)$ Ground distance $d_{ij} \geq 0$ Flow (from bin i to bin j) $g_{ij} \geq 0$ Potential π_i	Distributions $p_1(x), p_2(x)$ $\int p_1(x)dx = \int p_2(x)dx = 1$ Difference $p(x) := p_1(x) - p_2(x)$ Cost function $c(x,y) \geq 0$ Joint distribution $q(x,y) \geq 0$ Potential $f(x)$
EMD $:= \min \sum_{ij} g_{ij} d_{ij}$ s.t. $\sum_i g_{ik} - \sum_j g_{kj} = f(k)$	EMD $:= \inf \int c(x,y) q(x,y) dxdy$ s.t. $\int q(u,y)dy - \int q(x,u)dx = p(u)$
Dual EMD $:= \max \sum_i \pi_i f(i)$ s.t. $\pi_i - \pi_j \leq d_{ij}$	Dual EMD $:= \sup \int f(x)p(x)dx$ s.t. $f(x) - f(y) \leq c(x,y)$

Dual metrics and EMD

*C*onsider images I_i to be sensed by correlation with a collection of sensors f, in a convex set B.

*W*e can define a distance $\mathrm{d}_{B^*}(\mathrm{I}_i, \mathrm{I}_j) = \sup_{f \in B} \int_X f(x)(\mathrm{I}_i(x) - \mathrm{I}_j(x))dx$

If B is the unit ball in Holder classes we get the EMD distances ,
The point being that if B transforms nicely under certain distortions so does the dual metric.
The computation of the dual norm for standard classes of smoothness is linear in the number of samples. (Unlike the conventional EMD optimization or minimal distortion metrics)
This is applicable to general data sets , such as documents, or profiles .
Morever since dual norms are usually weighted combinations of l^p norms at different scales, it is easy to adjust the weights to account for noisy conditions. (ie redefining smoothness).

$$\mathrm{d}_{B^*}(\mathrm{I}_i,\mathrm{I}_j)=\sup_{f\in B}\int_X f(x)(\mathrm{I}_i(x)-\mathrm{I}_j(x))$$

if B is the unit ball in Holder α, then its obvious that the dual distance transforms well under small perturbation of the identity h(x)=x+r(x) , where r<ε

In fact

Let D(x)=I$(h(x))h'-\mathrm{I}(x)$

Then $\sup_{f\in B}\int_X f(x)\mathrm{I}(h(x))h'-\mathrm{I}(x)dx=\sup_{f\in B}\int_X I(x)(f(h(x))-f(x))dx<\varepsilon^{\alpha}$

This argument extends trivially to other metrics, dual to spaces in which changes of variables perform small perturbations in L^{∞}.

Unlike the direct EMD distance the dual distance , which is the dual norm of a Holder or Lipshitz class can easily be computed in a variety of ways , each of which has been proposed as a potential substitute for the EMD distance , and they all turn out to be equivalent .

The simplest way, starts with the observation that Holder functions are characterized by the boundedness of wavelet coefficients after rescaling so that the EMD corresponds to being integrable after dual rescaling . An equivalent definition is given by the sum over different scales of histograms.

A metric equivalent to Earth mover distance is obtained as follows

consider blurrred versions of the image at several scales

$$P_t(I)(x) = \int (1/t)\exp(|x-y|^2\ /t)I(y)dy$$

then

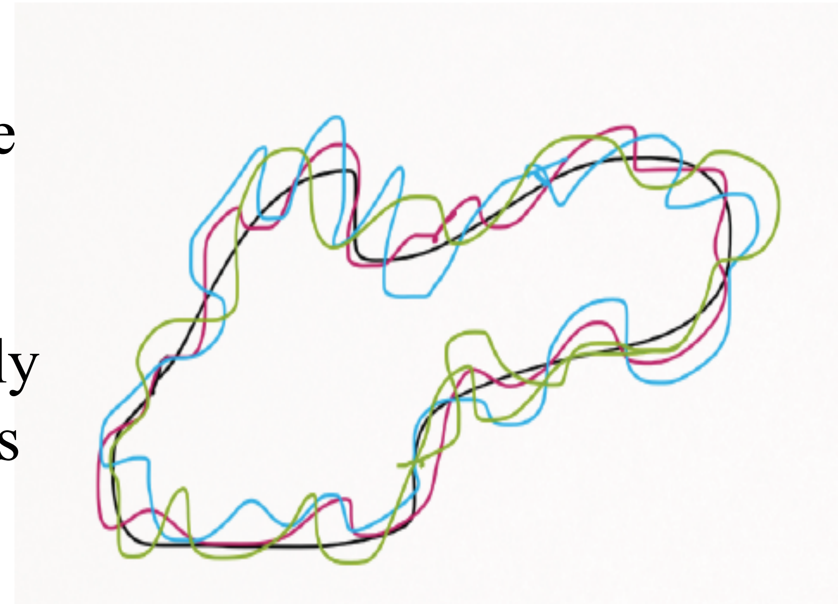
$$\mathrm{d}_\alpha(I_1,I_2) = \int_0^\infty t^{\alpha-1}\Big(\int_{R^2} |P_t(I_1-I_2)(x)|\,dx\Big)dt < \int_{R^2xR^2} \mathrm{d}_\alpha(x,y)\,|I_1-I_2|\,dxdy$$

is equivalent to EMD with distance Penalty $|x\text{-}y|^{\alpha 2} = c\mathrm{d}_\alpha(x,y)$.

$$\mathrm{d}_\alpha(x,y) = \int_0^\infty t^{\alpha-1}\Big(\int |P_t(x,u) - P_t(y,u)|\,du\Big)dt$$

If P_t is a more general diffusion process the same results hold.

Measuring distance between curves , becomes an easy exercise , finding the Median curve is quite easy, it is also easy to find a distribution best approximating all of the curves , simply take the median of wavelet coefficients of all given curves.

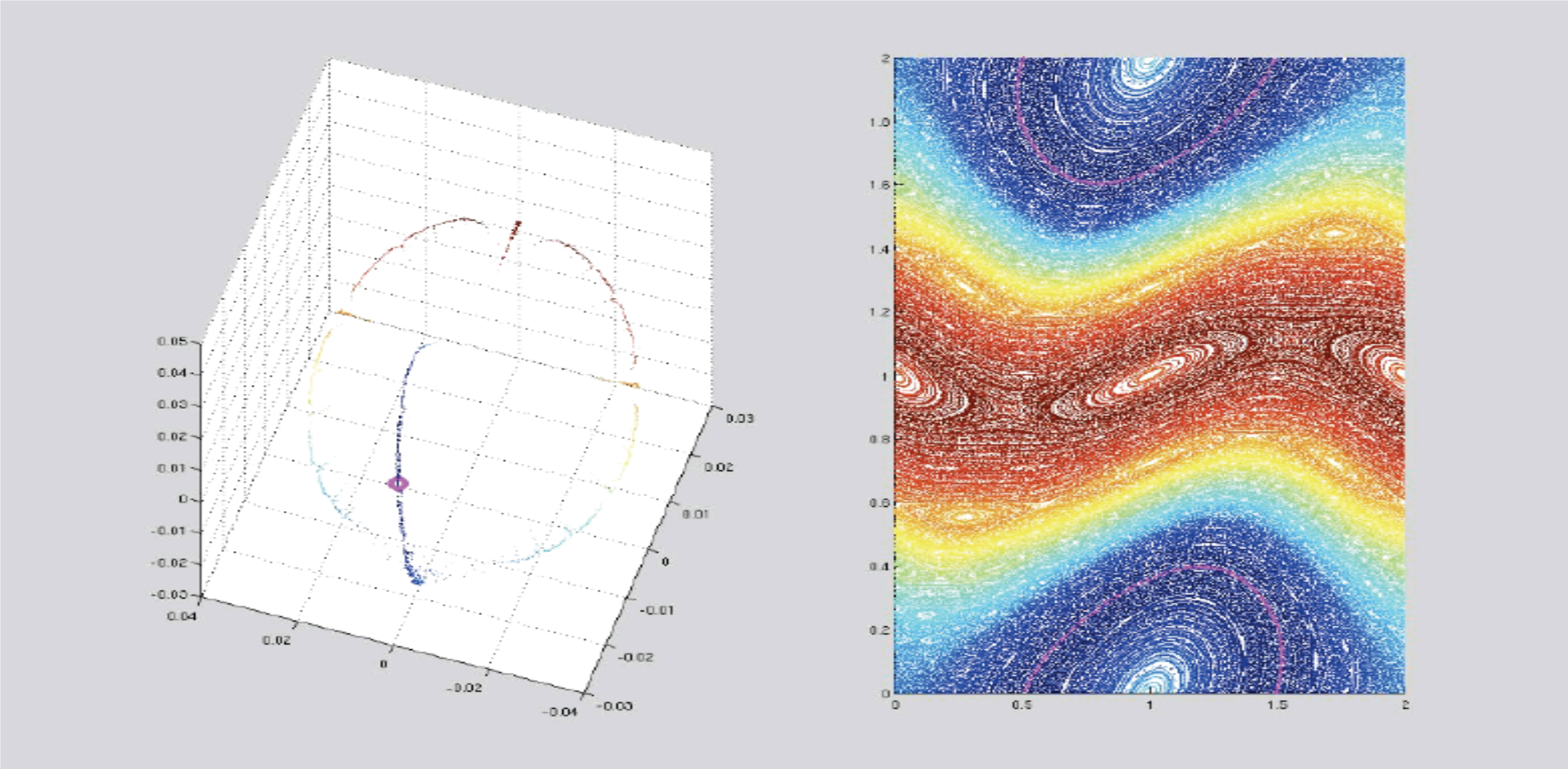


More generally this approach permits to build a transport between two probability measures , based on a multiscale histogram transport.

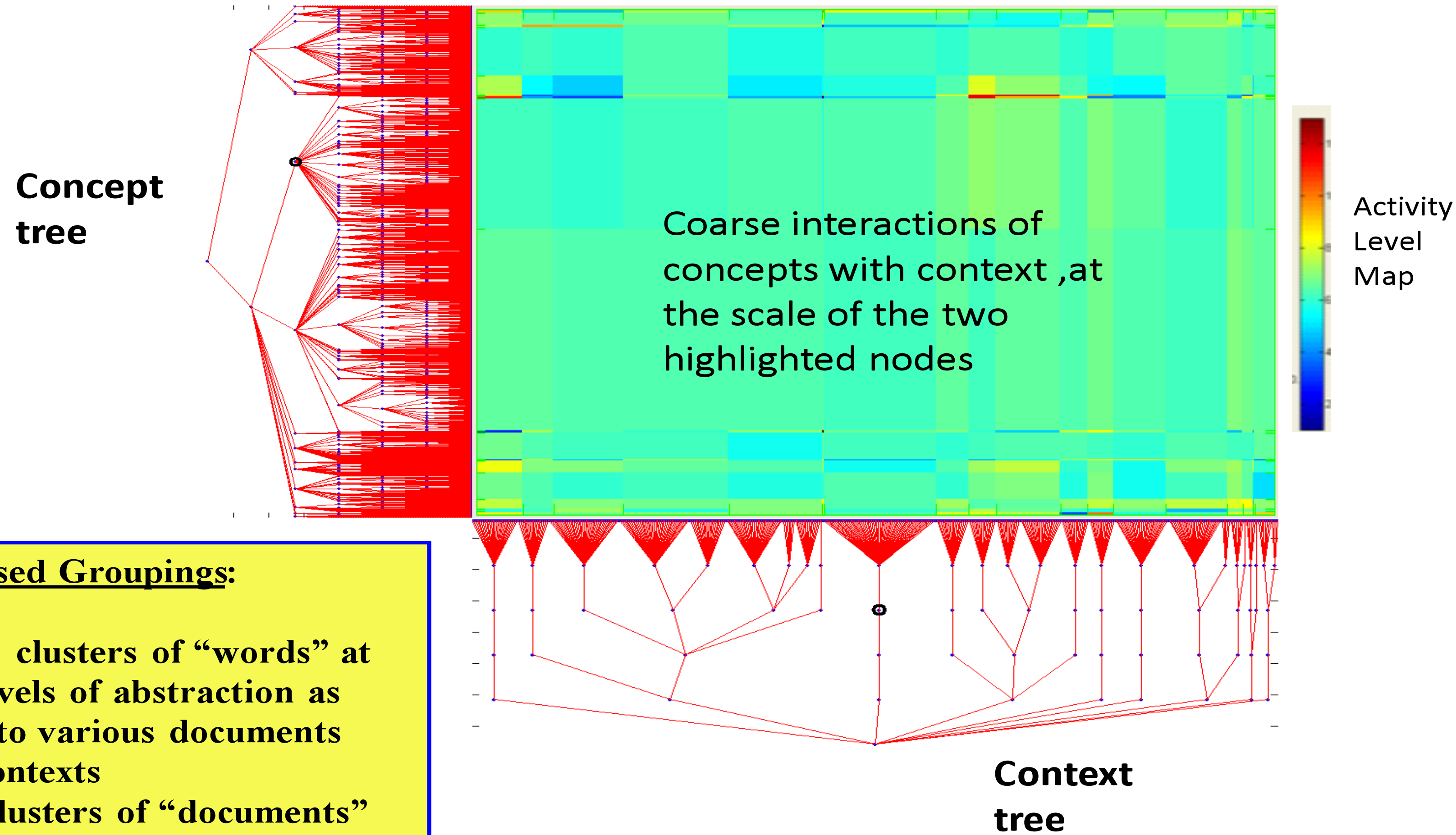
We now return to our original database analysis , in which both wavelet analysis and Besov spaces arise naturally, and where both emd And dual bi-holder distances arise naturally

Diffusion embedding of the graph of orbits of the standard map on the torus, each orbit is a measure , we use the earth moving distance to define distances between orbits and organize in a graph.

$$p_{\ell+1} \triangleq p_\ell + \alpha \sin(\theta_\ell),$$
$$\theta_{\ell+1} \triangleq \theta_\ell + p_{\ell+1},$$



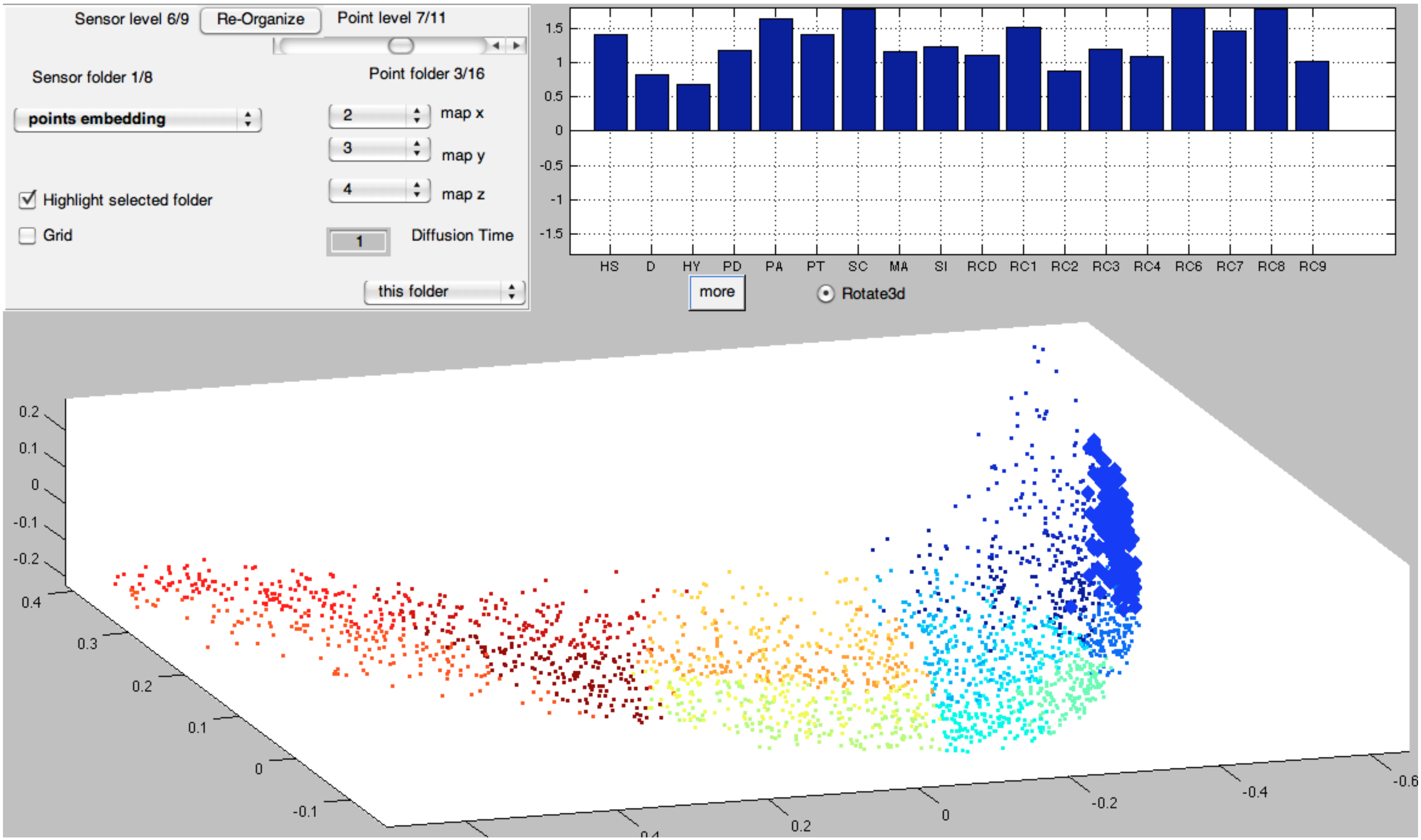
Mutual Organization / Tree Structures for context- concept duality,
Although we use linguistic analogies these trees were built on time series of observations of 500 objects , the concepts are scenarios of times with similar responses among the population while the contexts are group of objects with similar temporal responses.



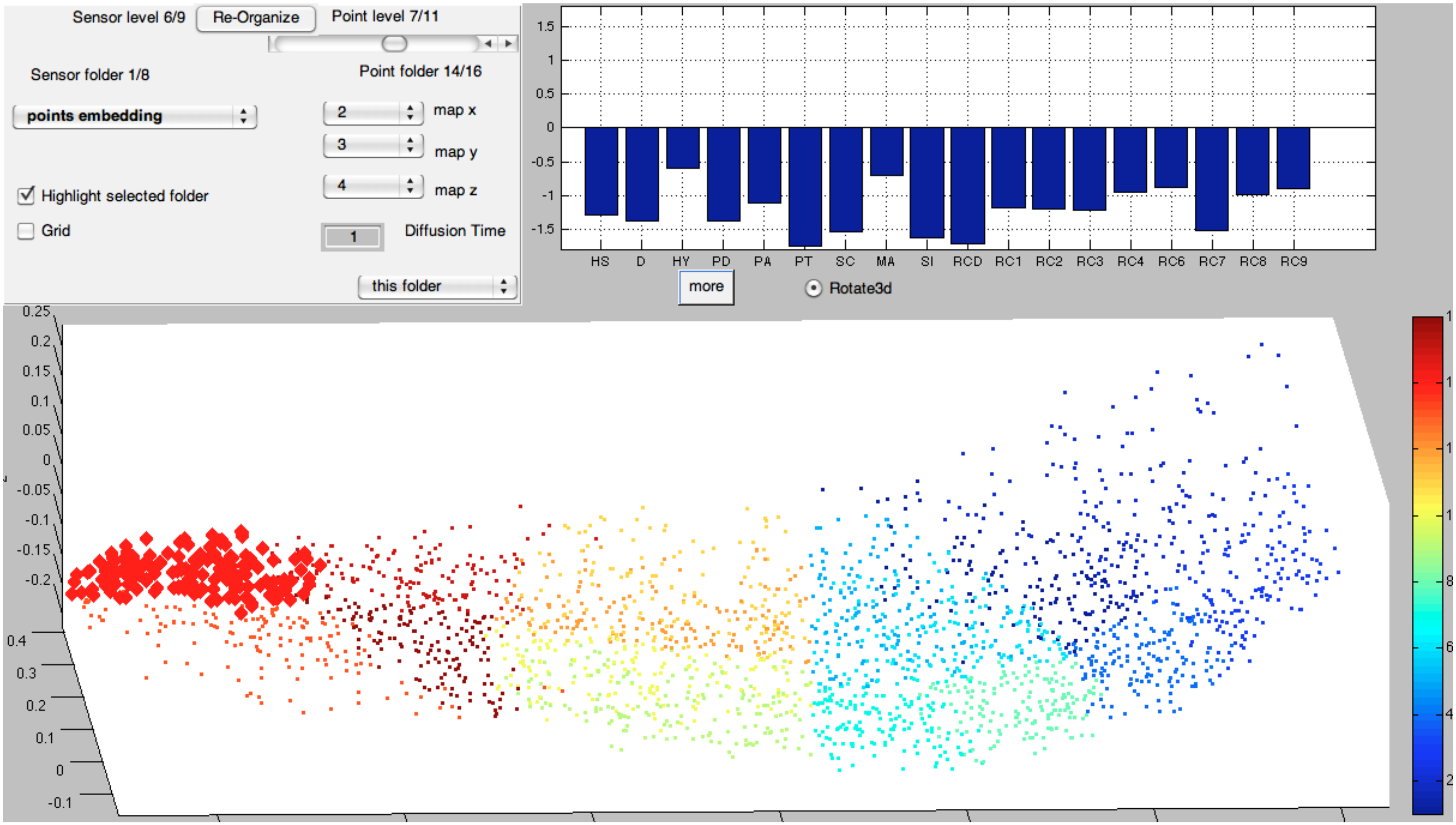
<u>Affinity based Groupings</u>:

Concepts - clusters of "words" at different levels of abstraction as they relate to various documents clusters =contexts
Contexts- clusters of "documents" with similar vocabulary profile

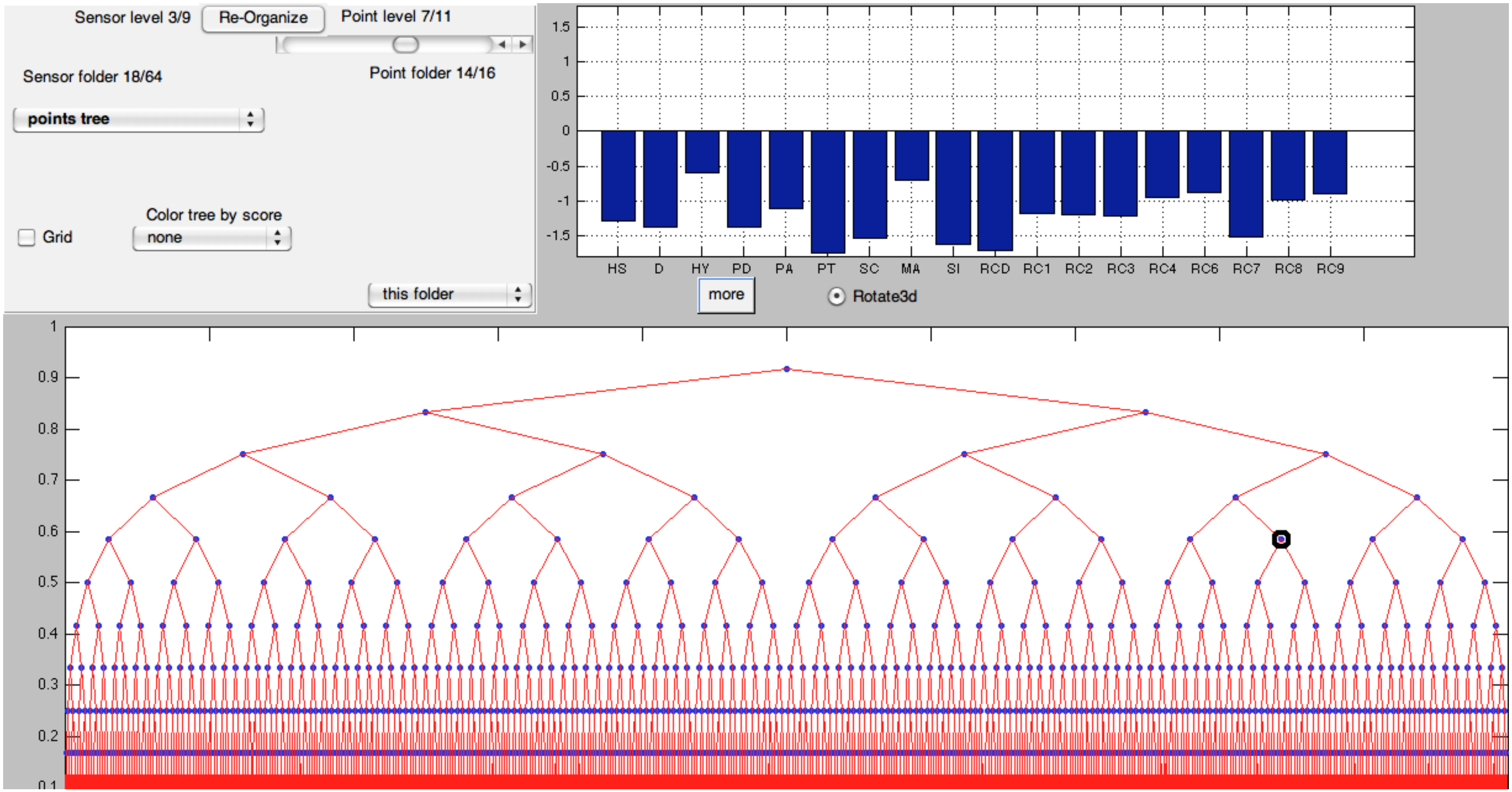
Demographic organization by earth mover distance among profiles of the population.
The blue highlighted group is on one extremity ,having problems.



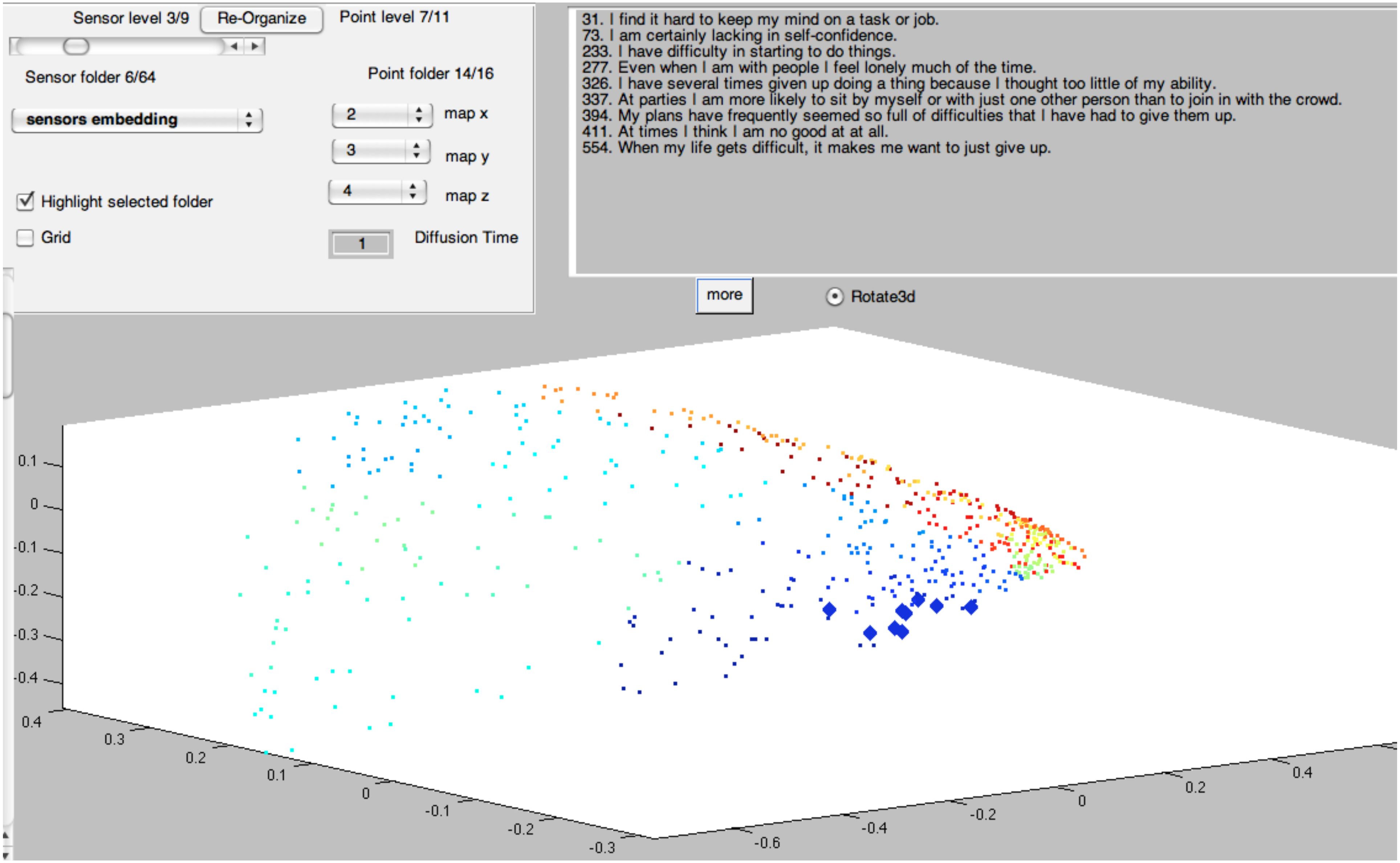
The red group is on the other end , being quite healthy .



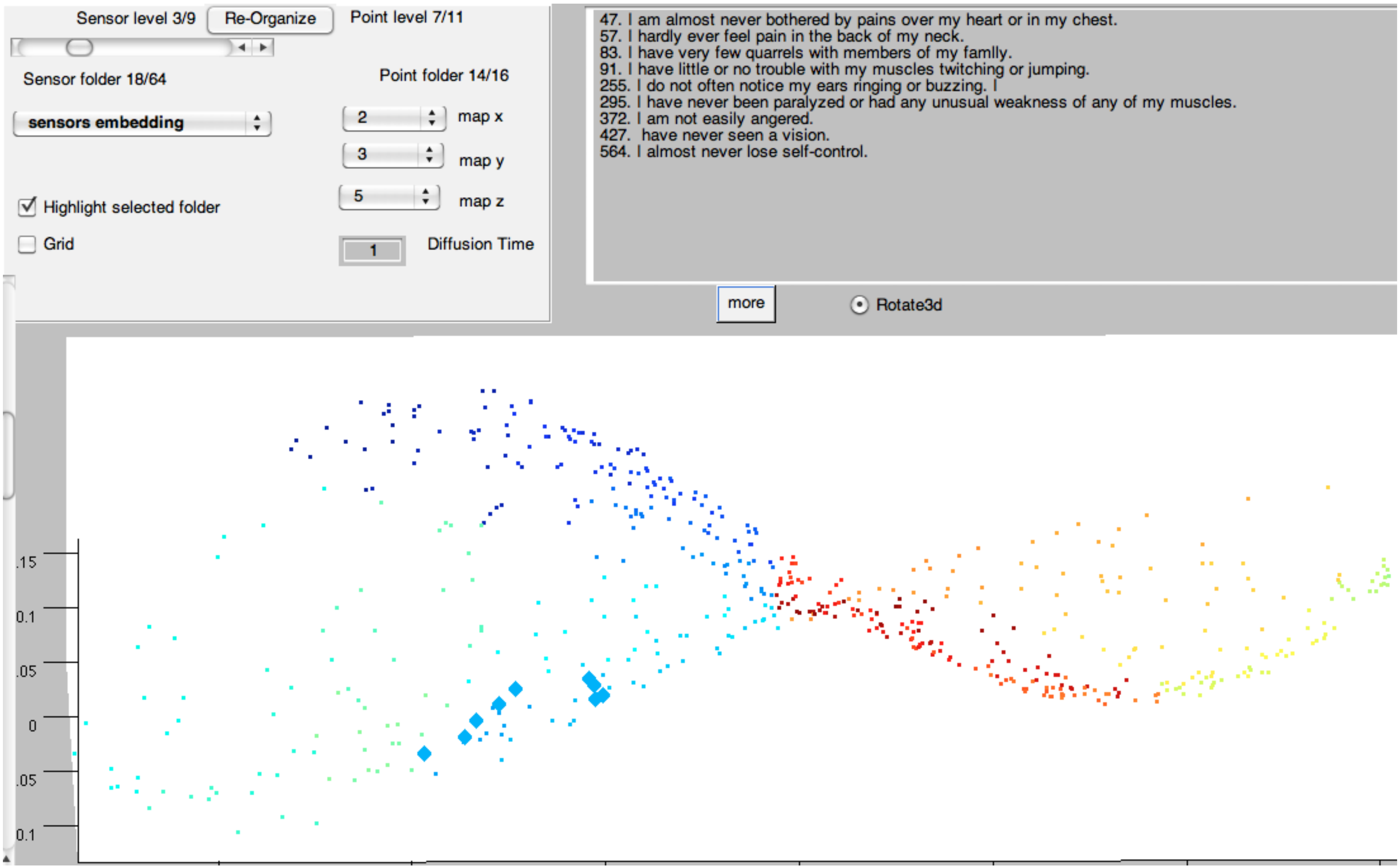
The demographic tree , where the previous red group is marked.



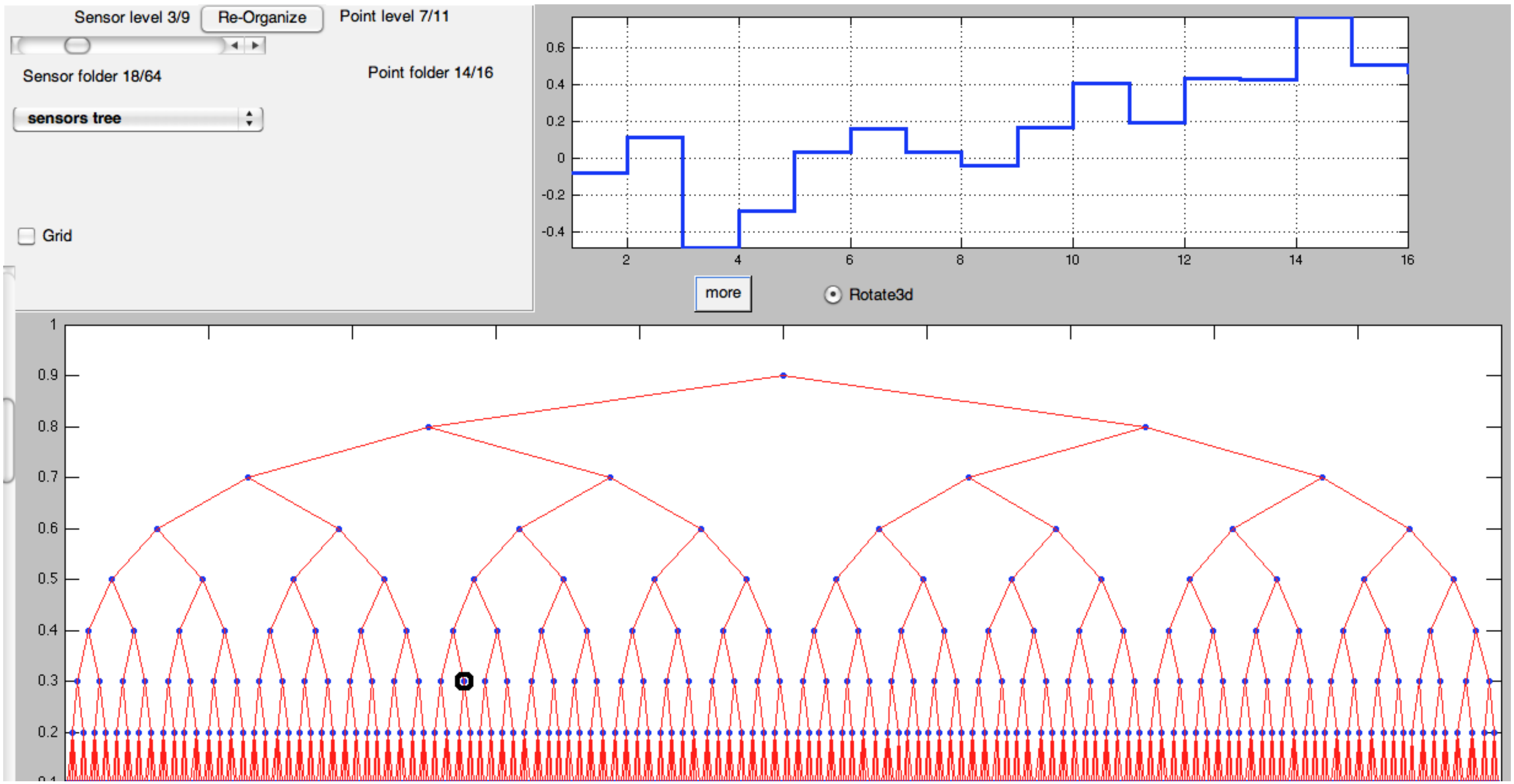
Conceptual organization of the questions into a geometry .



Another group of questions

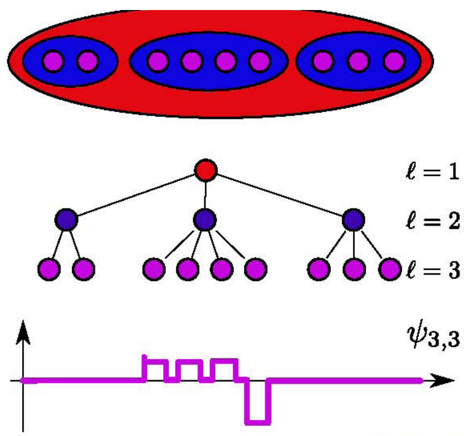


The same questions as above on the metaquestion tree , and the response profile of various demographic groups , on the left problem groups , on the right healthy people.

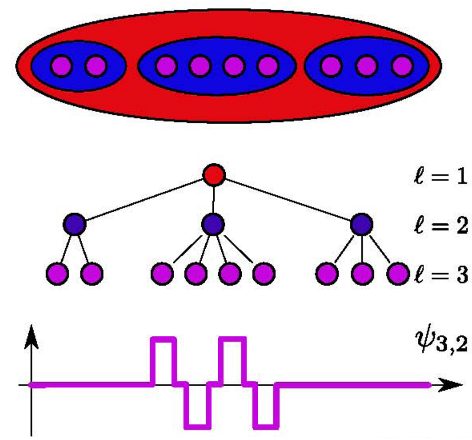


Observe that whenever we have a partition of data into a tree of subsets, we can associate with the tree an orthonormal basis constructed by orthogonalization of the characteristic functions of subsets of a parent node, first to the parent, and then to each other, as seen below.

This is precisely the construction of Haar wavelets on the binary tree of dyadic intervals or on a quadtree of dyadic squares .

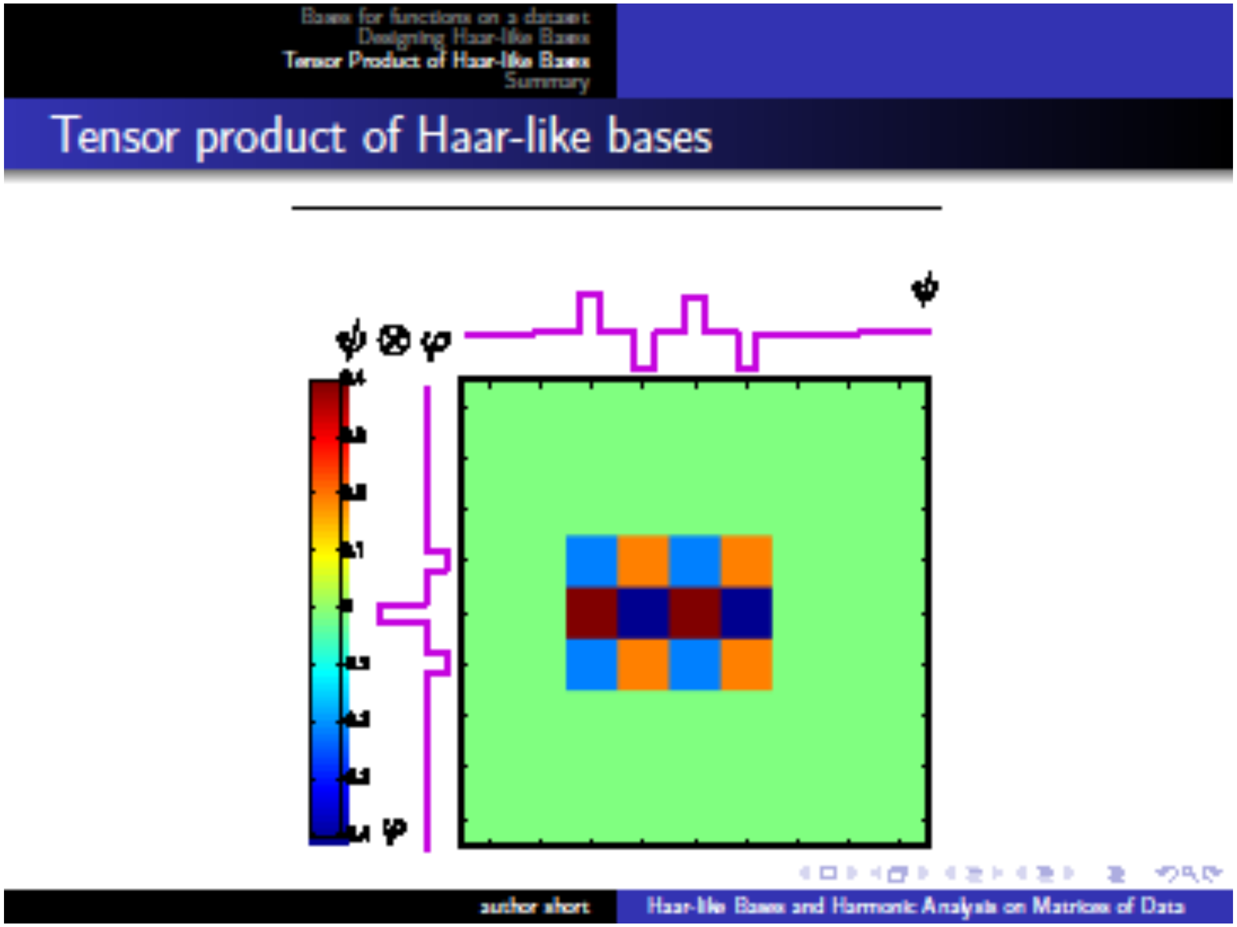


$d(x,y)^{\alpha}$



To a partition tree we associate a metric, which is the weight of the lowest folder containing two points , and of course we have corresponding notion of Holder regularity as well as an earth mover distance. Conversely any metric d(x,y) has the property that:

$d(x,y)^{\alpha}$ is the average of a small number of tree metrics for any $\alpha < 1$.



The tensor product basis indexed by bi-folders, or rectangles in the data base is used to expand the full data base .

The geometry is iterated until we can no longer reduce the entropy of the tensor-Haar expansion of the data base.

Definition 1. A *coherent matrix organization* of M is a pair of nontrivial metrics ρ_X on X and ρ_Y on it Y, such that for all $x_0, x_1 \in X$ and $y_0, y_1 \in Y$,

$$M(x_1, y_1) = M(x_1, y_0) + \left[M(x_0, y_1) - M(x_0, y_0)\right] + \epsilon$$

where $|\epsilon| < C \cdot \rho_X(x_0, x_1)^\alpha \cdot \rho_Y(y_0, y_1)^\alpha$ for constants $C > 0$ and $0 < \alpha \leq 1$. (The approximation error ϵ may depend on x_0, x_1, y_0, y_1).

In other words, in a coherent matrix organization, the value $f(x_1, y_1)$ can be estimated from entries at three neighboring points with quadratic (to the α) error. This condition is simply the Taylor expansion form of the so-called Mixed-Hölder condition

$$|M(x_0, y_0) - M(x_0, y_1) - M(x_1, y_0) + M(x_1, y_1)| \leq C \cdot \rho_X(x_0, x_1)^\alpha \cdot \rho_Y(y_0, y_1)^\alpha$$

A basic analytical observation on Haar like Basis functions is that a natural Entropy condition such as

$$\Sigma | a_R | < 1$$

on the coefficients of an expansion does not only enable sparse representations but also implies smoothness as well as accuracy of representation , in a ***dimensionally independent*** estimate with number of terms $< \quad 1/\varepsilon$

observe that $\quad | h_R(x) | \le \dfrac{\chi_R(x)}{|R|^{1/2}}$ and

$$\int \left| f - \sum_{R>\varepsilon} a_R h_R(x) \right| dx = \int \left| \sum_{R\le\varepsilon} a_R h_R(x) \right| dx < \int \sum_{R\le\varepsilon} | a_R | \frac{\chi_R(x)}{|R|^{1/2}} dx < \varepsilon^{1/2} \Sigma \left| a_R \right| \ ,$$

$(\beta = 0)$

Moreover $\quad \displaystyle\int \left| f - \sum_{|R|>\varepsilon, |a_R|>\varepsilon} a_R h_R(x) \right| dx < \varepsilon^{1/2}$

The entropy condition for standard wavelet basis in d dimensions corresponds to having d/2 derivatives in the (special atom) Hardy space H^1

Given a tree of subsets we can define a natural distance ρ(x,y)

as the size of the smalles folder (node) containining the two points ,

we say that a function is Holder of order β if

$|f(x)-f(y)| < c\rho(x,y)^{\beta}$ (or its variance on any folder F $< c|F|^{\beta}$)

this condition is equivalent to the following condition

on the Haar coefficients

$$\left|a_R\right| < c\left|R\right|^{1/2+\beta}$$

We claim that if f satisfies the condition $\sum\left|a_R\right| < 1$ then it is locally Holder of order1/2

More precisely there is a decreasing sequence of sets E_l such that $|E_l| \leq 2^{-l}$

and a decomposition (of Calderon Zygmund type)

$f = g_l + b_l$ where b_l is supported

on E_l . and g_l is Holder β=1/2 with constant $2^{(l+1)}$

or equivalently with Haar coefficients satisfying $\left|a_R\right| < 2^{(l+1)}\left|R\right|^{1/2+1/2}$

All of this, extends to tensor products for the Bi Holder case, with R=IxJ.

Observe that in reality there is no need to build a Haar system it suffices to consider the matingale differences and the corresponding Besov spaces ie.

let E_l be the conditional expectation on the partition at level l

and $\Delta_l = E_{l+1} - E_l$, clearly we have

f=$\sum(E_{l+1} - E_l)f \quad + E_0 f$, the entropy condition is the equivalent

to

$\int \sum_l |\Delta_l(f)\, 2^{l/2}| \quad <\infty$, i,e 1/2 a derivative in L^1.

$\int \sum_l |\Delta_l(f)\, 2^{-l/2}|$ is the dual norm to Holder of index 1/2 equivalent to the emd

with that distance.

References

[1] R.Coifman ,M Gavish Geometric Analysis of Databases and Matrices Applied and Computational Harmonic Analysis 2012.

[2] R. Coifman and G. Weiss, Analyse Harmonique Noncommutative sur Certains Espaces Homogenes, Springer-Verlag, 1971.}

[3] R. Coifman ,G. Weiss, Extensions of Hardy spaces and their use in analysis. *Bul. Of the A.M.S.,* **83**, #**4**, 1977, 569-645.

[4] Belkin, M., & Niyogi, P. (2001). Laplacian eigenmaps and spectral techniques for embedding and clustering. Advances in Neural Information Processing Systems 14 (NIPS 2001) (p. 585).

[5]Belkin, M., & Niyogi, P. (2003a). Laplacian eigenmaps for dimensionality reduction and data repre-
sentation. Neural Computation, 6, 1373{1396.

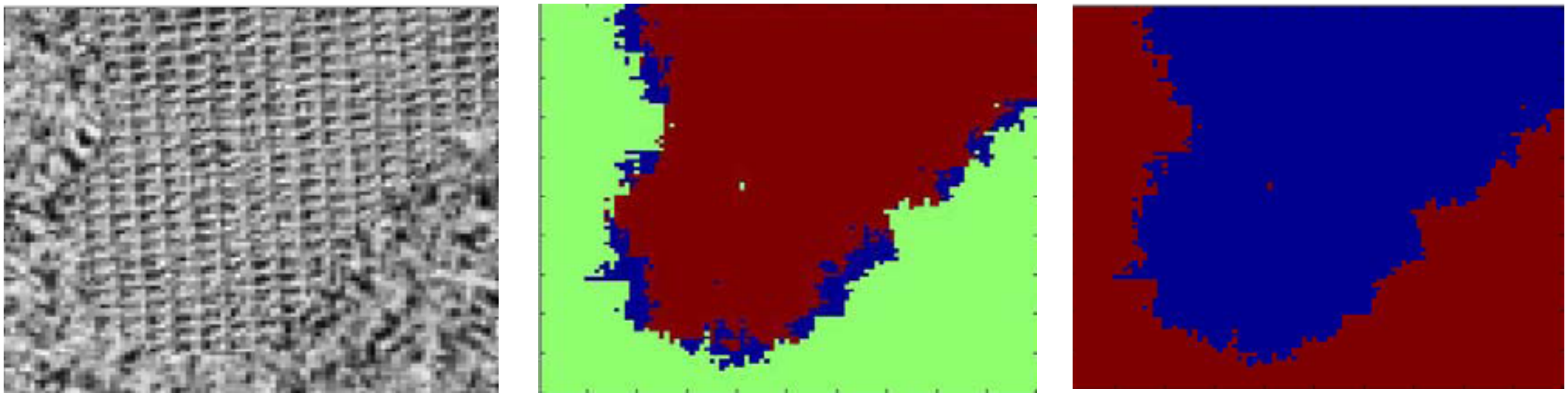
[6]Coifman, R. R., Lafon, S., Lee, A., Maggioni, M.,Nadler, B., Warner, F., & Zucker, S. (2005a) . Geometric diffusions as a tool for harmonic analysis and structure defnition of data. part i: Diffusion maps.Proc. of Nat. Acad. Sci., 7426{7431.

[7] Coifman R.R.,S Lafon, Diffusion maps, *Applied and Computational Harmonic Analysis,* 21: 5-30, 2006.

[8] Coifman R.R., B.Nadler, S Lafon, I G Kevrekidis, Diffusion maps, spectral clustering and reaction coordinates of dynamical systems, *Applied and Computational Harmonic Analysis,* 21:113-127, 2006.

[9] Ronald R Coifman1, Mauro Maggioni1, Steven W Zucker1 and Ioannis G Kevrekidis "Geometric diffusions for the analysis of data from sensor networks" Current Opinion in Neurobiology 2005, 15:576–584

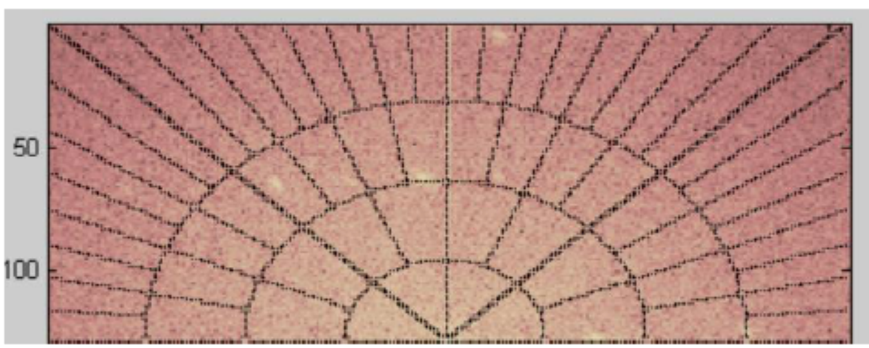
[10] R Coifman W. Leeb Earth Mover's Distance and Equivalent Metrics for Hierarchical Partition Trees, Yale CS technical report July 2013.

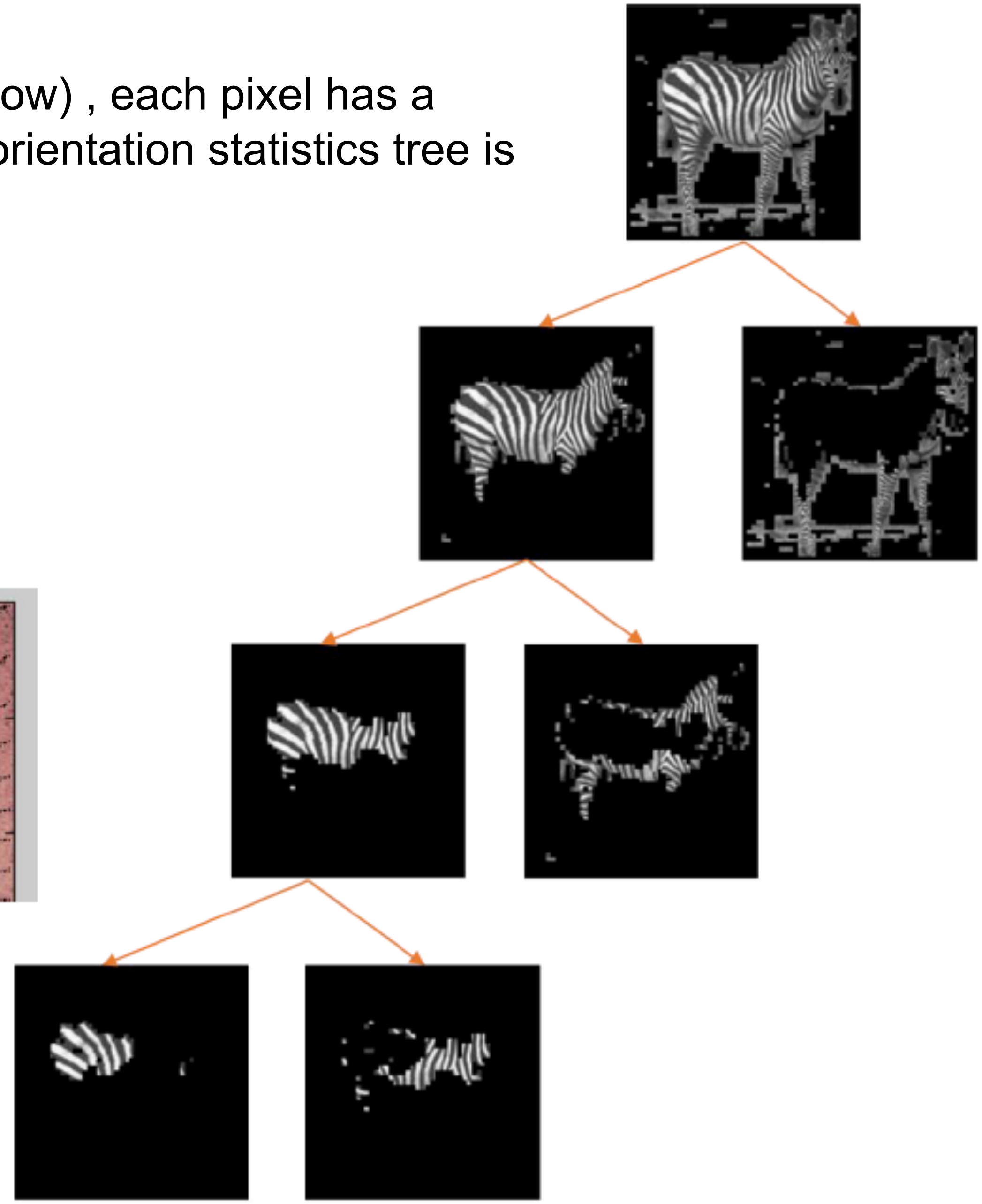


The same approach of organizing an image as a questionnaire ,is effective for texture segmentation.
Here we associate with each pixel the log values of the fourier coefficients of the 11X11 square centered at the pixel .
The middle image shows folders at a level before last ,observe the spot in the middle of the brown .
The image on the right is a good segmentation of the textures .
Observe that no assumptions or filters were given , this can be done as easily without using the FT.

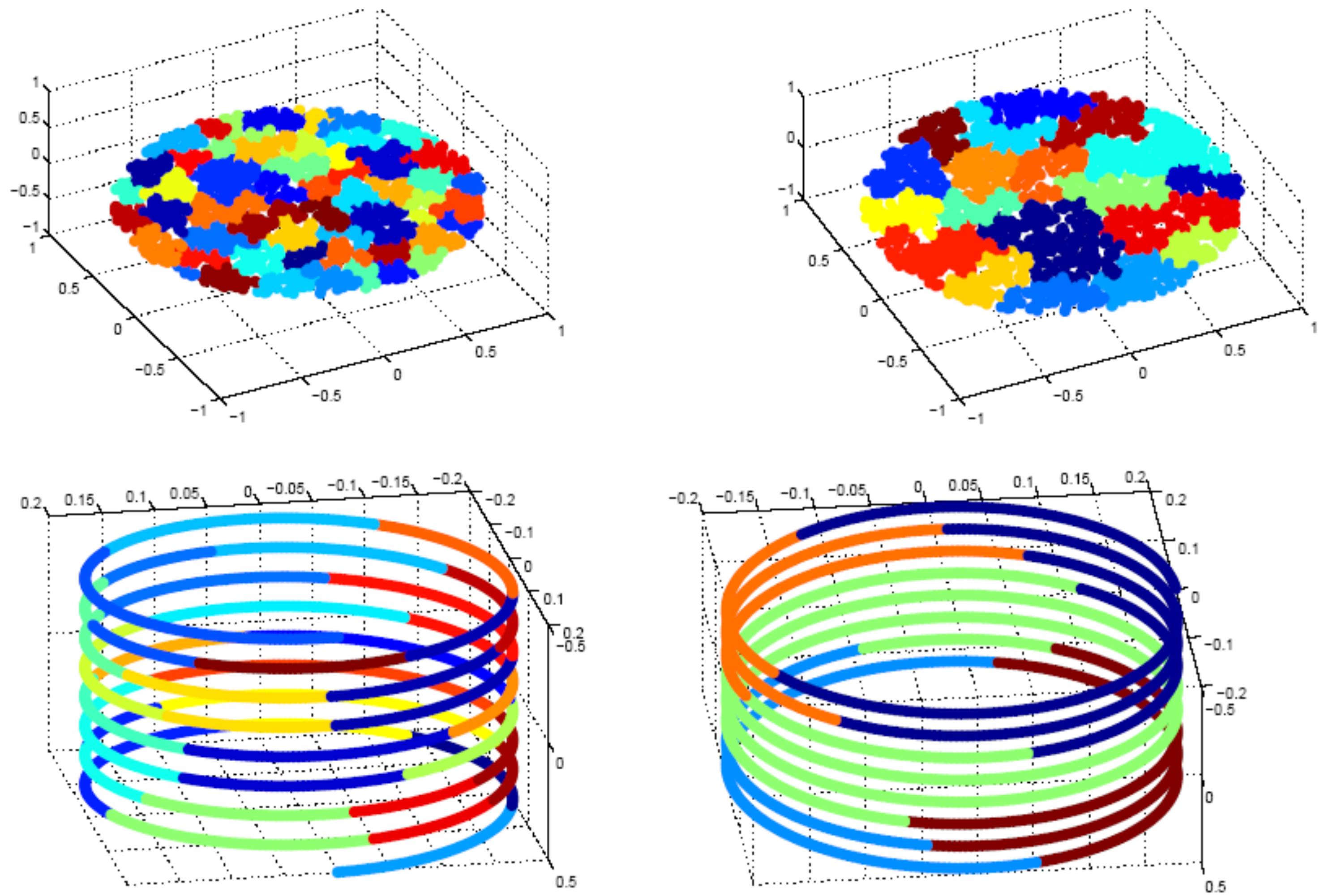
Questioning the Zebra image.

We use about 60 band pass filters (below) , each pixel has a response to a given filter . The stripe orientation statistics tree is generated





One of the first applications of wavelet bases ,was the observation that CZ operators could be efficiently implemented in such bases .
Assume more generally that we have the matrix of potentials of a collection of sources located on a spiral, which are evaluated on a flat disk located away . We need to find a wavelet basis on each structure relative to its geometry. The full matrix is then expanded efficiently in the Tensor Wavelet basis .
Observe also that a matrix is usually given in garbled order.



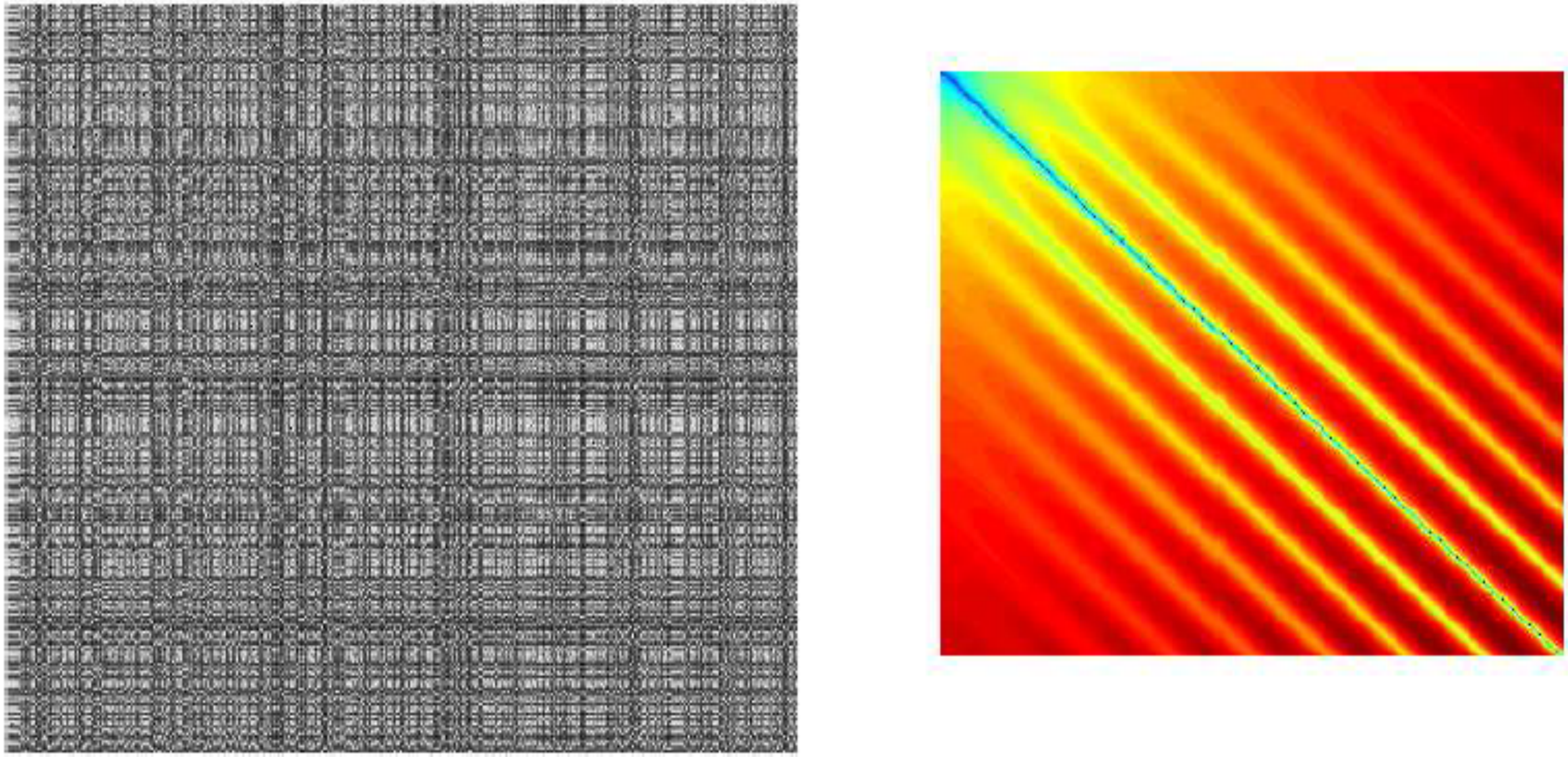


Figure 5.7: The kernel $||x-y||^{1/4}$ on the spiral, before and after permutation.

A simple empirical diffusion 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 a discrete Graph Laplace Operator.

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

$$X_i \to X_i^{(t)} = (\lambda_1^{\,t}\varphi_1(X_i), \lambda_2^{\,t}\varphi_2(X_i), \lambda_3^{\,t}\varphi_3(X_i), ..)$$

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

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

Another similar construction for empirical data

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.

Observe that in general any positive kernel with spectrum as above can give rise to a natural orthogonal basis as well as a natural multiscale analysis.

Let k be a positive definite kernel whose restriction to the data set is expanded in eigenfunctions

$$k(x,y)=\sum \lambda_i^2 \varphi_i(x)\varphi_i(y)$$

Let

$$D^2(x,y)=\sum \lambda_i^2 (\varphi_i(x)-\varphi_i(y))^2$$

Then

$$k(x,x)+k(y,y)-2k(x,y)=D^2(x,y)$$

Clearly D is a distance on the data induced by the Geometric short time Diffusion map

$$\mathrm{x}\in\Gamma\rightarrow \widehat{\mathrm{X}}^{\mathrm{t}}(\mathrm{x})=\{\lambda_i^t\varphi_i(x)\}\in \mathrm{l}^2 \ .$$

.

The multiscale tree building organization algorithm proceeds as follows .

Start with a disjoint partition of the graph into clusters of diameter between 1 and 2 relative to the distance at scale 1 .

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.

On this graph we partition again into clusters of diameter between 1 and 2 relative to the set distance (we double the time scale) and redefine the affinity between clusters of clusters using the previously defined affinity between sub clusters.

Iterate until only disjoint clusters are left.
Another approximate version of this algorithm is to embed the data using a diffusion map into Euclidean space and pull back a Euclidean based version of the above .

Learning and extrapolating functions.

A simple method to tune the geometry at it relates to various queries or functional approximation is obtained as follows.

Start with a function known on a subset of the data , and find a simple/ smooth function agreeing with it , for example a Haar expansion with minimal norm in l^1 , use that function as a last row= question with an appropriate weight to reorganize the questionnaire geometry as it relates to that question , and iterate the process.

.

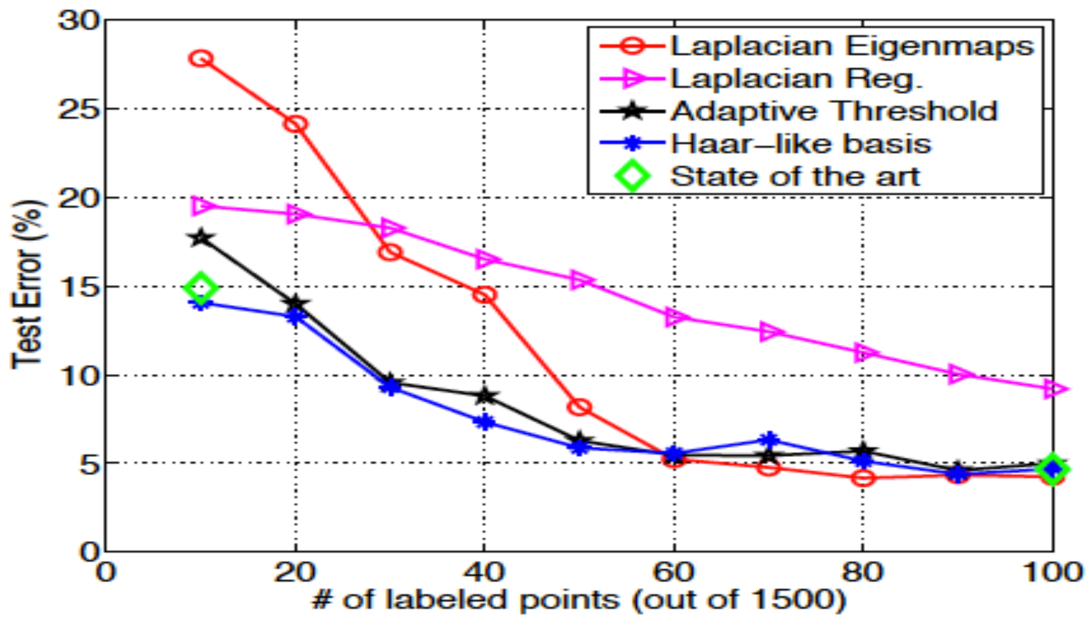


Figure 2. Results on the USPS benchmark.

Table 1. Test classification errors for USPS benchmark

Method	10 labeled	100 labeled
1-NN	19.82	7.64
SVM	20.03	9.75
MVU + 1-NN	14.88	6.09
LEM + 1-NN	19.14	6.09
QC + CMN	**13.61**	6.36
Discrete Reg.	16.07	**4.68**
TSVM	25.20	9.77
SGT	25.36	6.80
Cluster-Kernel	19.41	9.68
Data-Dep. Reg.	17.96	5.10
LDS	17.57	4.96
Laplacian RLS	18.99	**4.68**
CHM (normed)	20.53	7.65
Haar-like	**14.01**	**4.70**