Mathematical Foundations of Multiscale graph representations and interactive learning

Mauro Maggioni
Dept. of Mathematics and Computer Science
Duke University

FODAVA review meeting
12/3/2009

Joint work with E. Monson, R. Brady; P.W. Jones, A. Little, L. Rosasco, R. Schul
Partial support: NSF/DHS, ONR, Sloan, Duke

Data in High-D

A deluge of data: documents, web searching, customer databases, hyper-spectral imagery, social networks, gene arrays, proteomics data, sensor networks, financial transactions, traffic statistics (automobilistic, computer networks)...

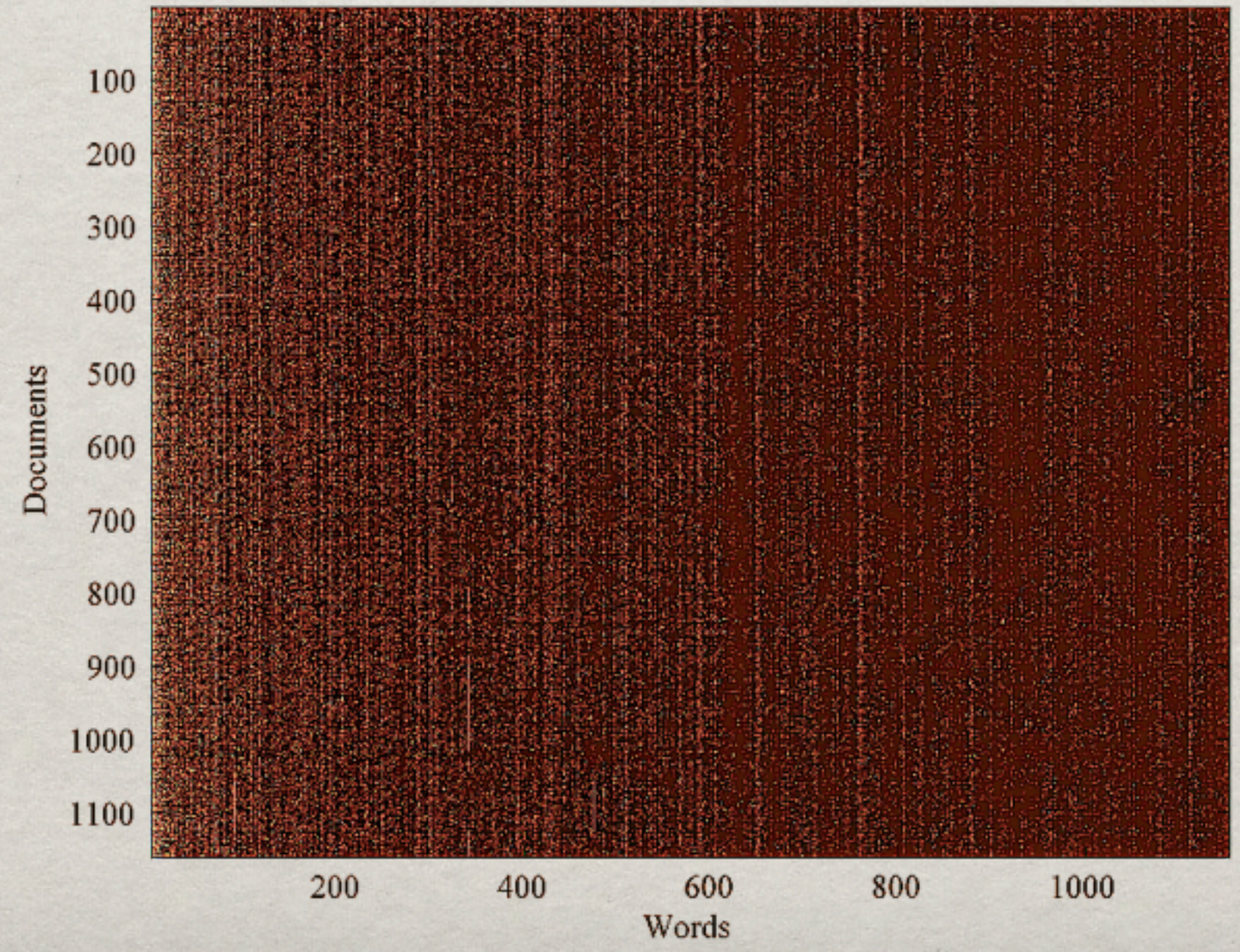
Common feature: data is given in a high dimensional space, however it has a much *lower dimensional intrinsic geometry*.
(i) *physical constraints*: for example the effective state-space of at least some proteins seems low-dimensional, at least when viewed at the time scale when important processes (e.g. folding) take place.
(ii) *statistical constraints*: for example many dependencies among word frequencies in a document corpus force the distribution of word frequency to low-dimensional, compared to the dimensionality of the whole space.

Example 1: Text documents

About 1100 Science News articles, from 8 different categories. We compute about 1000 coordinates, i-th coordinate of document d represents frequency in document d of the i-th word in a dictionary.

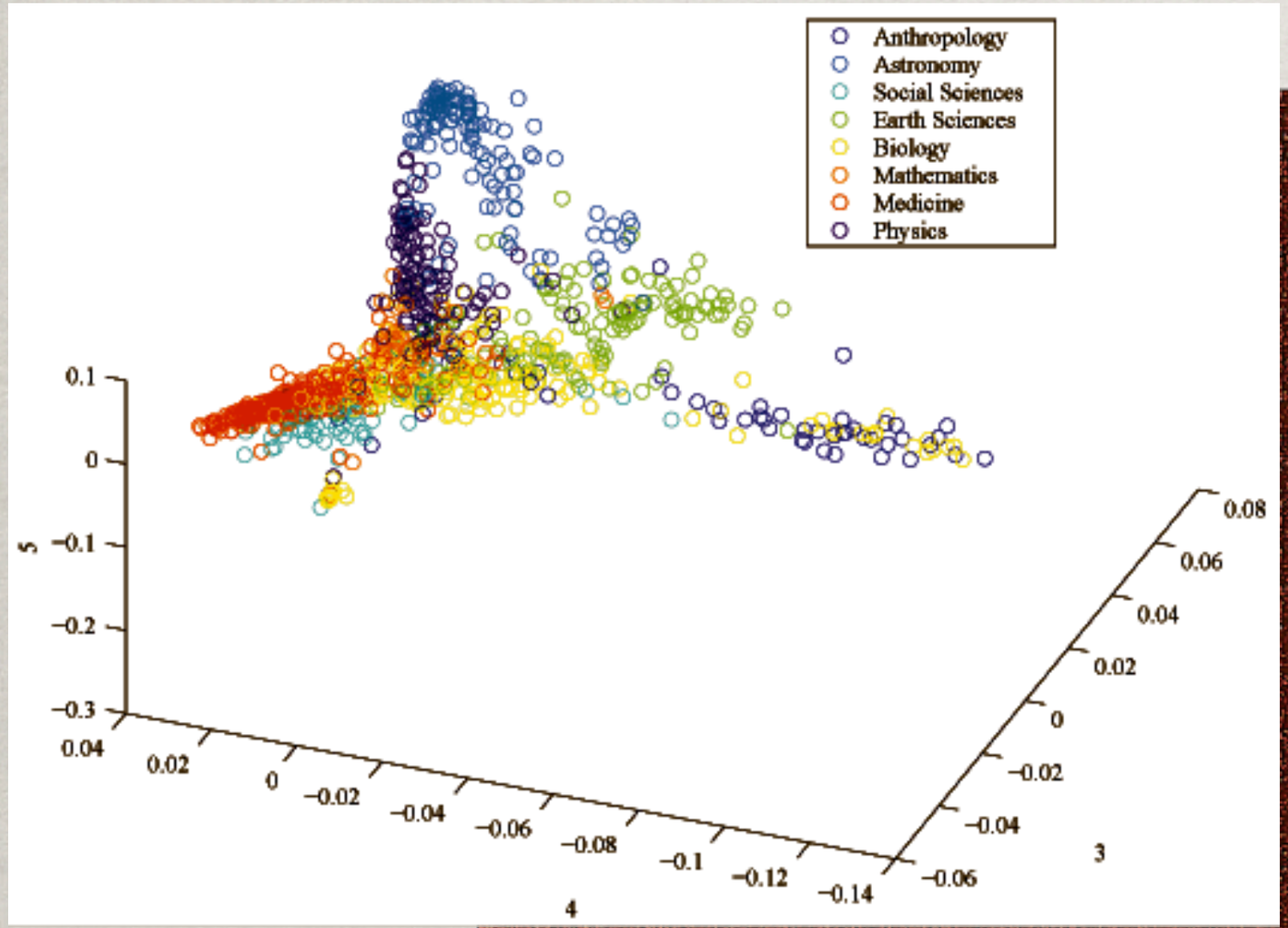
Example 1: Text documents

About 1100 Science News articles, from 8 different categories. We compute about 1000 coordinates, i-th coordinate of document d represents frequency in document d of the i-th word in a dictionary.



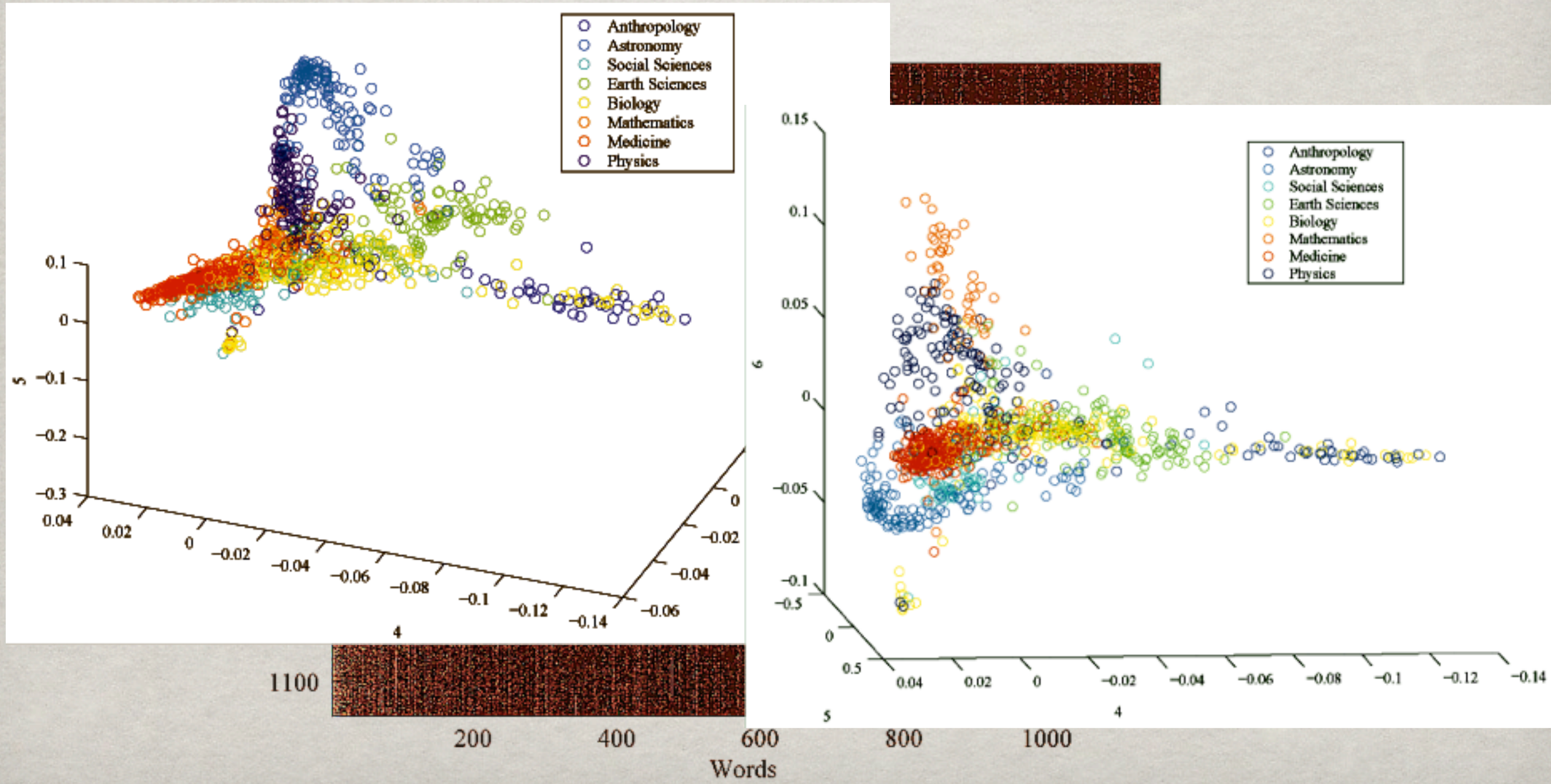
Example 1: Text documents

About 1100 Science News articles, from 8 different categories. We compute about 1000 coordinates, i-th coordinate of document d represents frequency in document d of the i-th word in a dictionary.



Example 1: Text documents

About 1100 Science News articles, from 8 different categories. We compute about 1000 coordinates, i-th coordinate of document d represents frequency in document d of the i-th word in a dictionary.

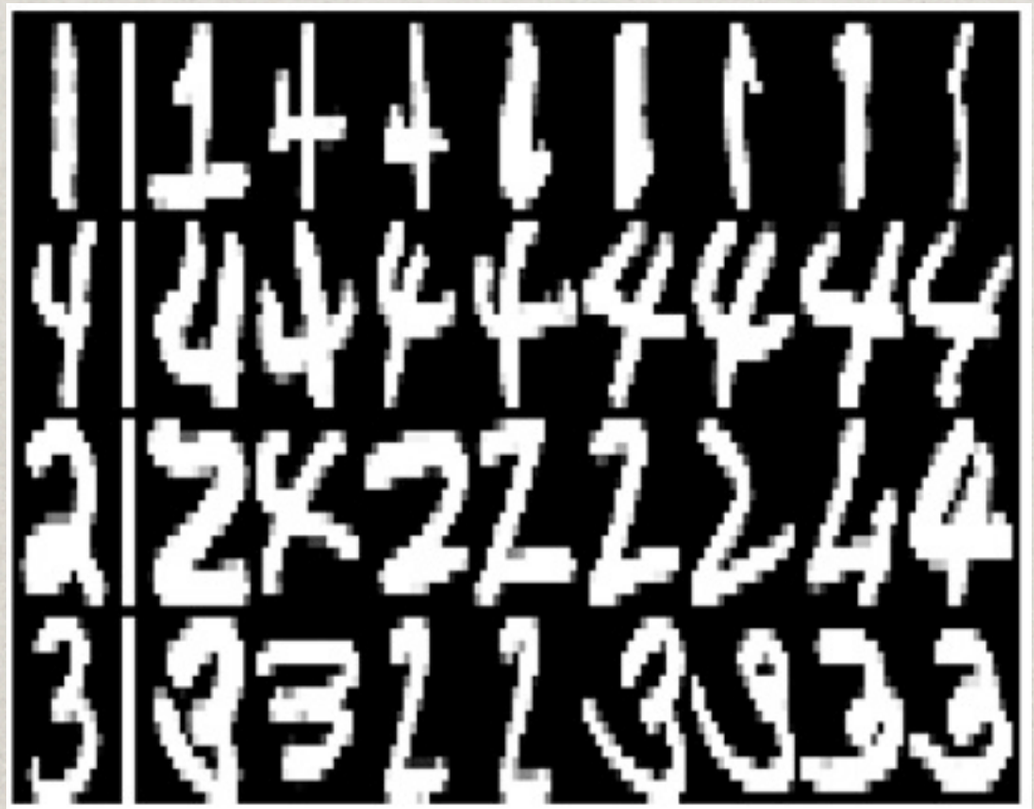


Example 2: Handwritten digits

Data base of about 60,000 28x28 gray-scale pictures of handwritten digits, collected by USPS. Point cloud in R^{728}.
Goal: automatic recognition.

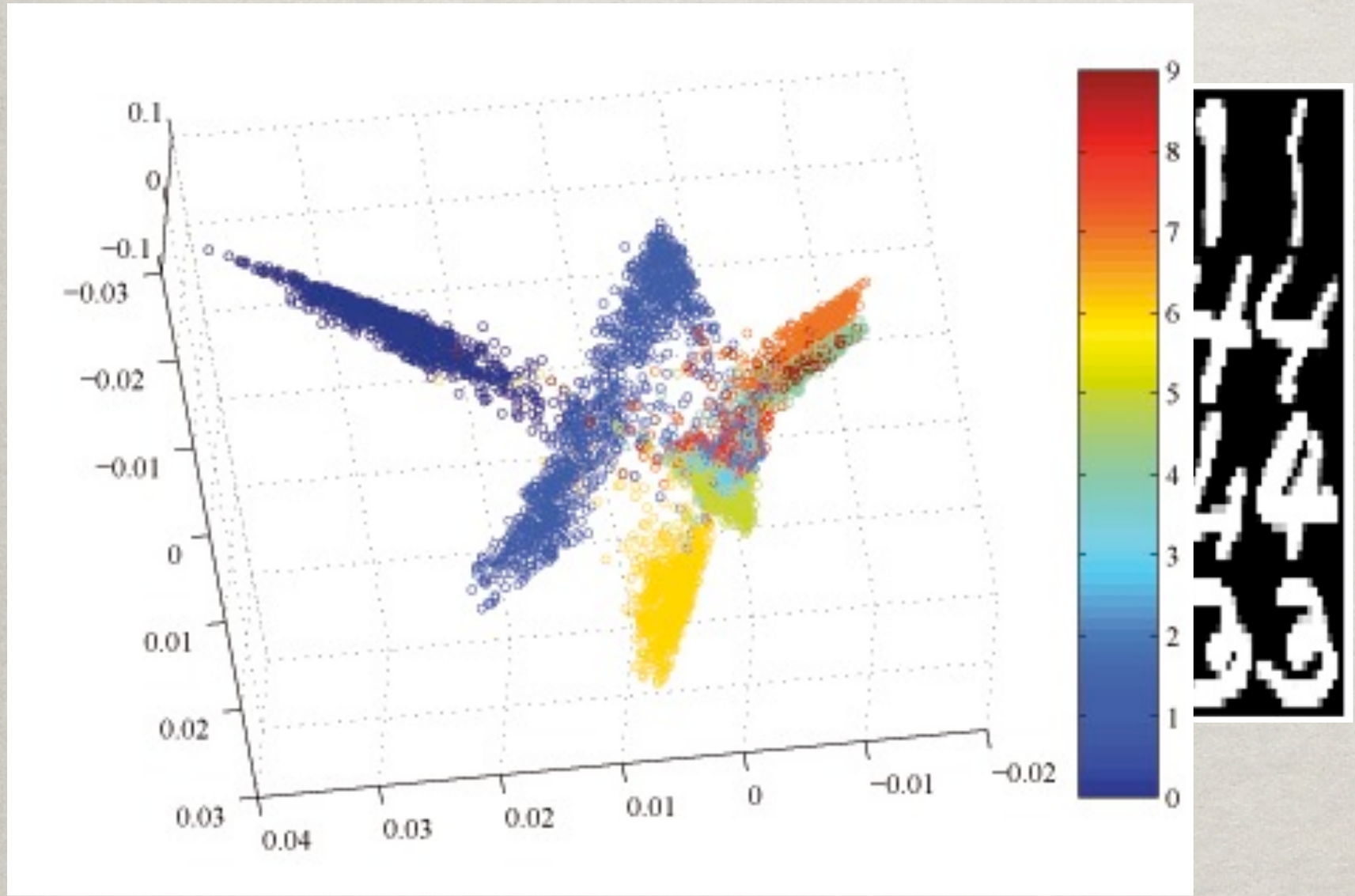
Example 2: Handwritten digits

Data base of about 60,000 28x28 gray-scale pictures of handwritten digits, collected by USPS. Point cloud in R^{728}.
Goal: automatic recognition.



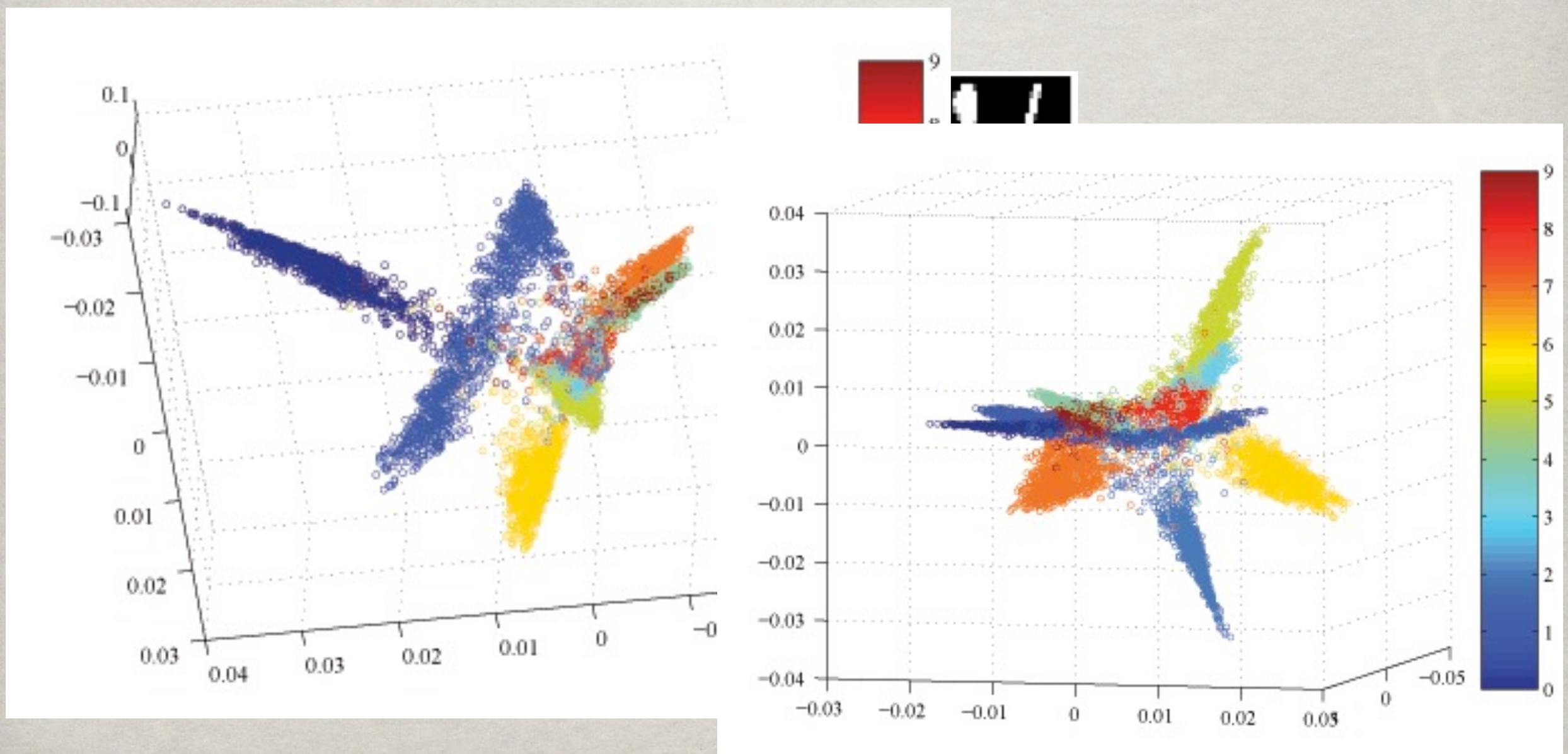
Example 2: Handwritten digits

Data base of about 60,000 28x28 gray-scale pictures of handwritten digits, collected by USPS. Point cloud in R^{728}.
Goal: automatic recognition.



Example 2: Handwritten digits

Data base of about 60,000 28x28 gray-scale pictures of handwritten digits, collected by USPS. Point cloud in R^{728}.
Goal: automatic recognition.



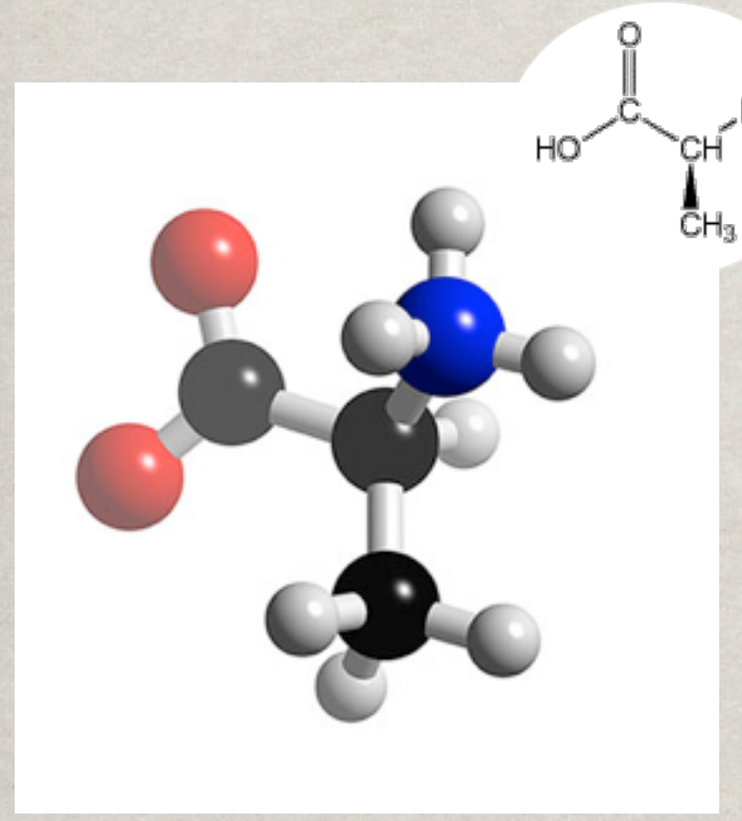
Example 3: Molecular

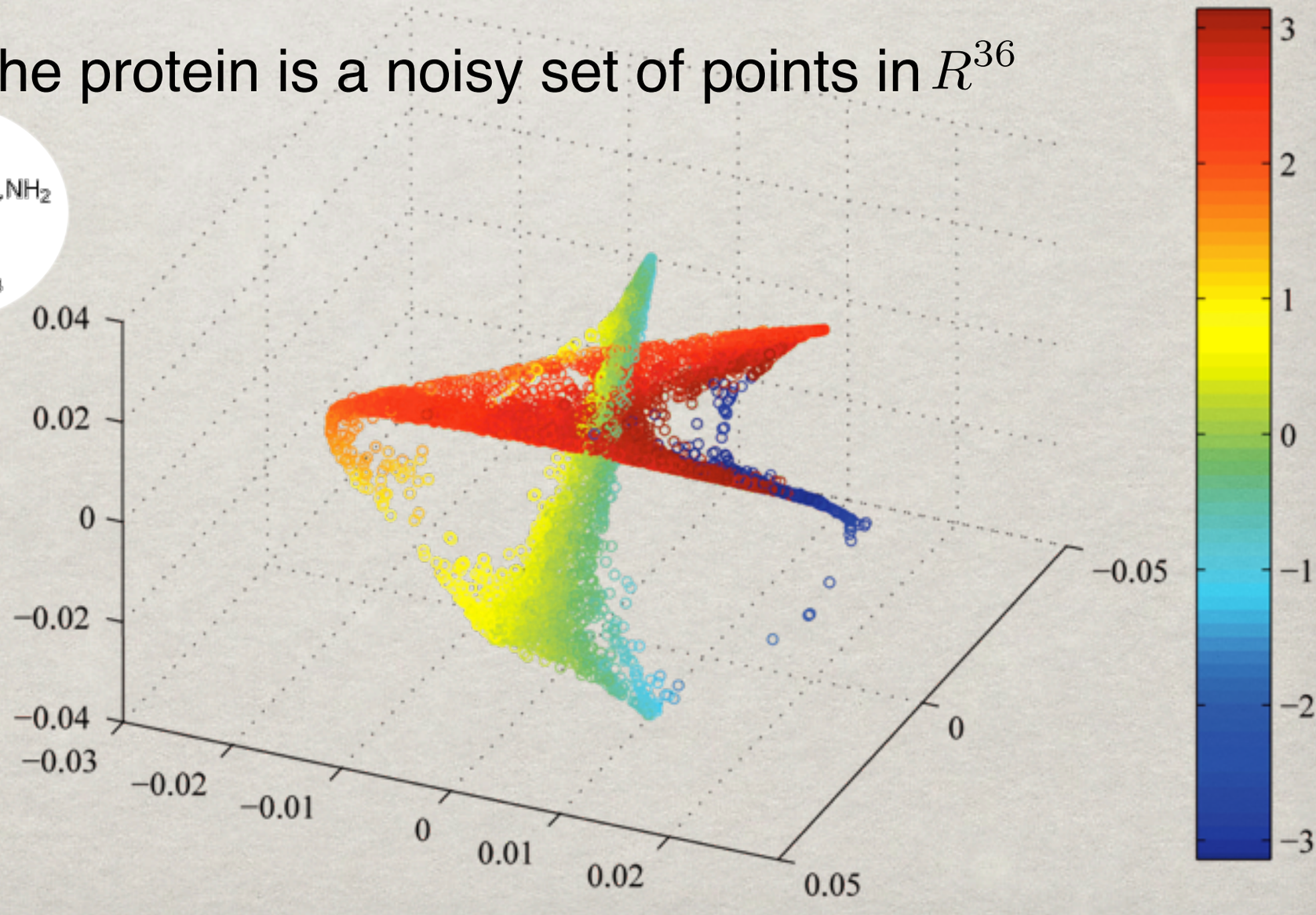
Joint with C. Clementi

The dynamics of a small protein (12 atoms, H atoms removed) in a bath of water molecules is approximated by a Langevin system of stochastic equations:

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

The set of states of the protein is a noisy set of points in R^{36}





Ongoing efforts in several directions

- Using diffusion processes on graphs for (inter)active learning.
- Perform multiscale analysis on graphs: construction of graph-adaptive multiscale analysis, for graph visualization and exploration, and (inter)active learning.
- Estimating intrinsic dimensionality of data
- Construct data-adaptive dictionaries for data-modeling and exploration.
- Apply recent results of provably good parametrizations of manifolds with heat kernels and eigenfunctions of the Laplacian

Random walks on data, Graphs

Given:

- Data $X = \{x_i\}_{i=1}^N \subset \mathbb{R}^D$.
- Local similarities via a kernel function $W(x_i, x_j) \geq 0$.

Simplest example: $W_\sigma(x_i, x_j) = e^{-||x_i - x_j||^2/\sigma}$.
Model the data as a weighted graph (G, E, W): vertices represent data points, edges connect x_i, x_j with weight $W_{ij} := W(x_i, x_j)$, when positive. Let $D_{ii} = \sum_j W_{ij}$ and

$$\underbrace{P = D^{-1}W}_{\text{random walk}}, \quad \underbrace{T = D^{-\frac{1}{2}}WD^{-\frac{1}{2}}}_{symm.\ ``random\ walk''}, \quad \underbrace{H = e^{-tL}}_{Heat\ kernel}$$

Here $L = I - T$ is the normalized Laplacian.

Note: W depends on the type of data. Moreover, W should be "local", i.e. close to 0 for points not sufficiently close.

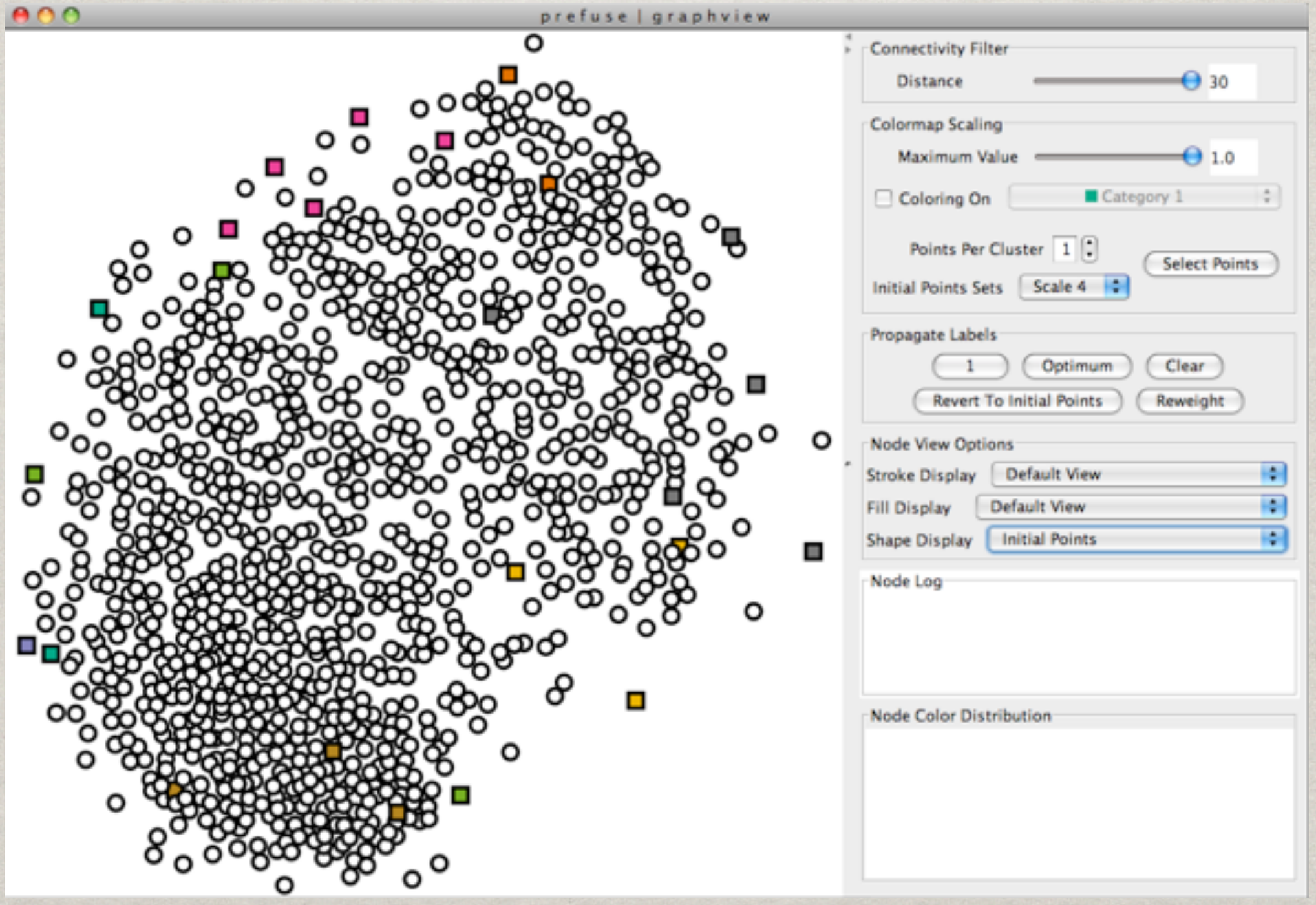
Some basic properties of r.w.'s

- $P^t(x, y)$ is the probability of jumping from x to y in t steps
- $P^t(x, \cdot)$ is a "probability bump" on the graph
- P and T are similar, therefore share the same eigenvalues $\{\lambda_i\}$ and the eigenfunctions are related by a simple transformation. Let $T\varphi_i = \lambda_i \varphi_i$, with $1 = \lambda_1 \geq \lambda_2 \geq \dots$.
- "typically" P (or T) is large and sparse, but its high powers are full and low-rank
- one can take limits as $n \to \infty$ of the above, when the points are sampled from a manifold $\mathcal{M}$, and recover in the limit natural operators such as Laplacian, heat kernels etc... on $\mathcal{M}$.

(Inter)Active Learning

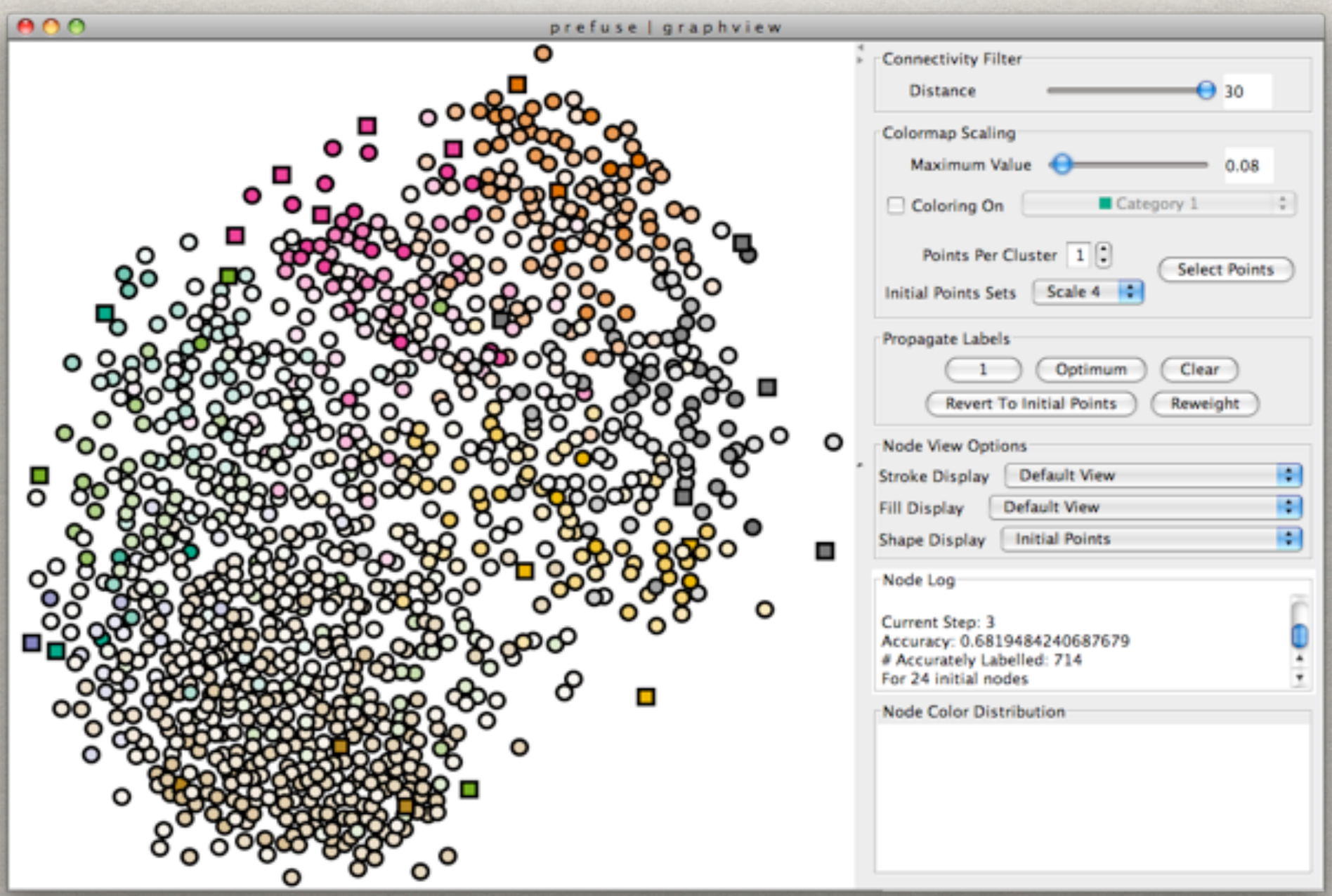
With E. Monson and R. Brady [C.S.]

A simple tool to explore possible algorithms



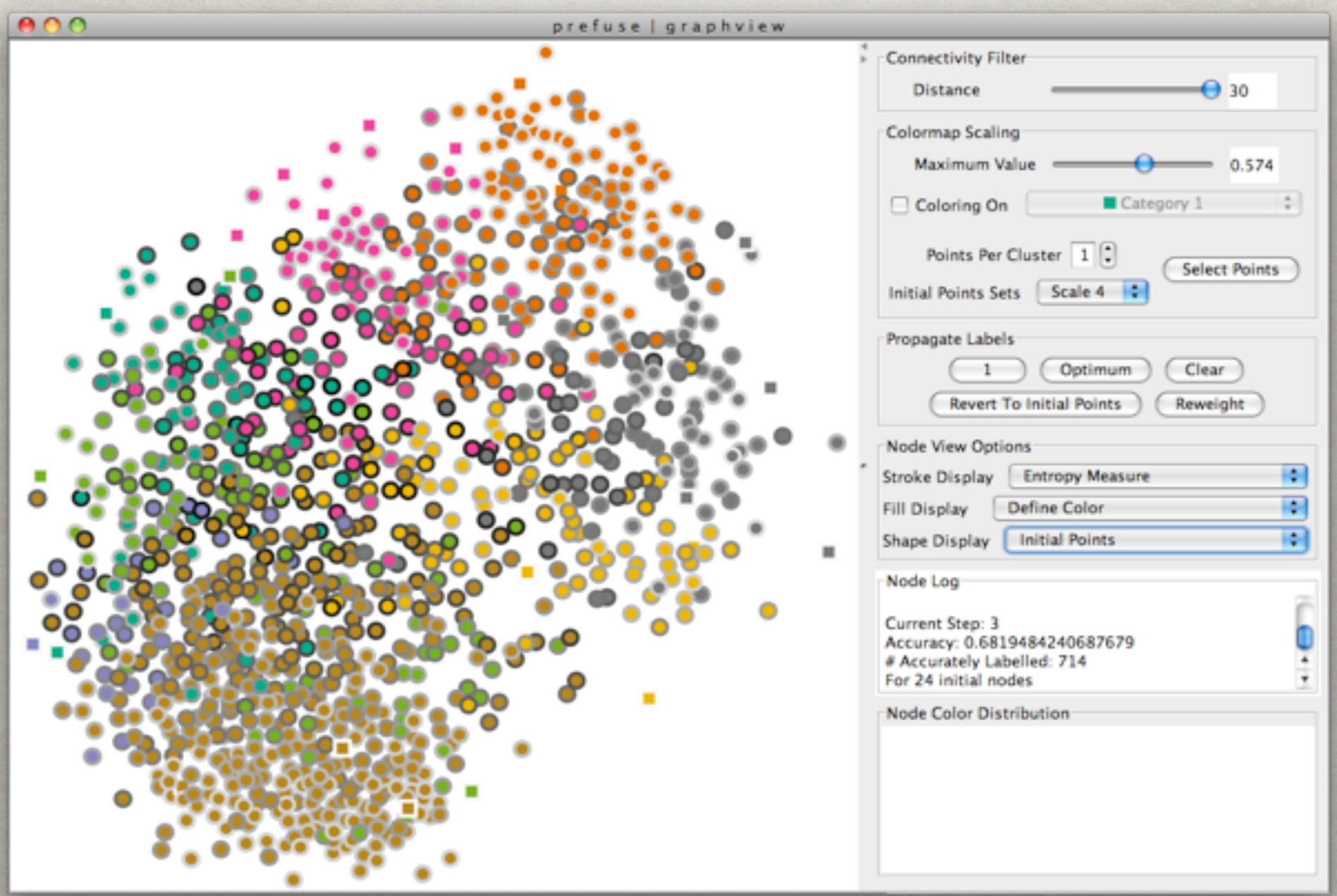
(Inter)Active Learning

With E. Monson and R. Brady [C.S.]



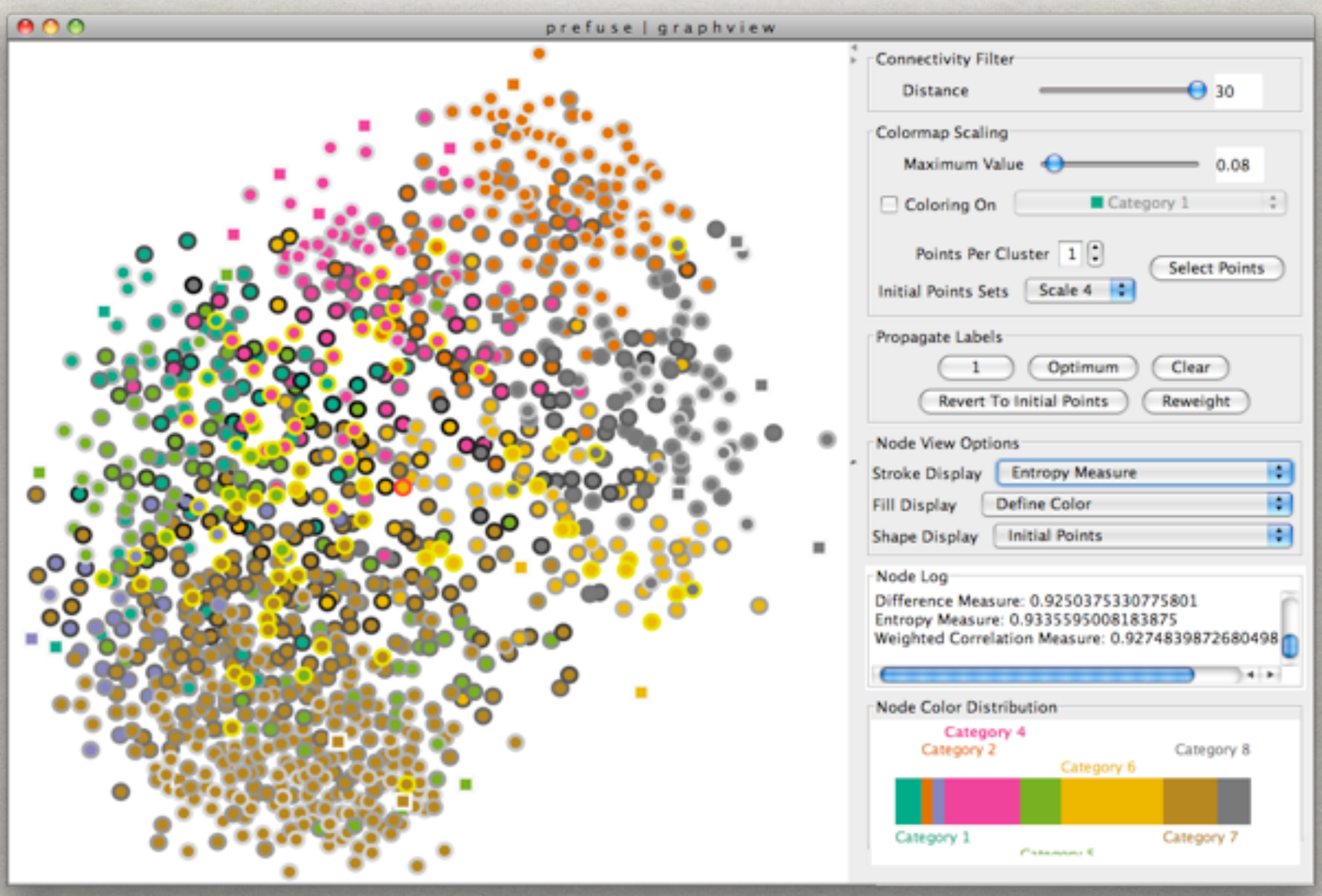
(Inter)Active Learning

With E. Monson and R. Brady [C.S.]



(Inter)Active Learning

With E. Monson and R. Brady [C.S.]



Graph is actually changed as labels are added - can update its visualization!

Multiscale Analysis on Graphs

With R. Coifman [Math.]

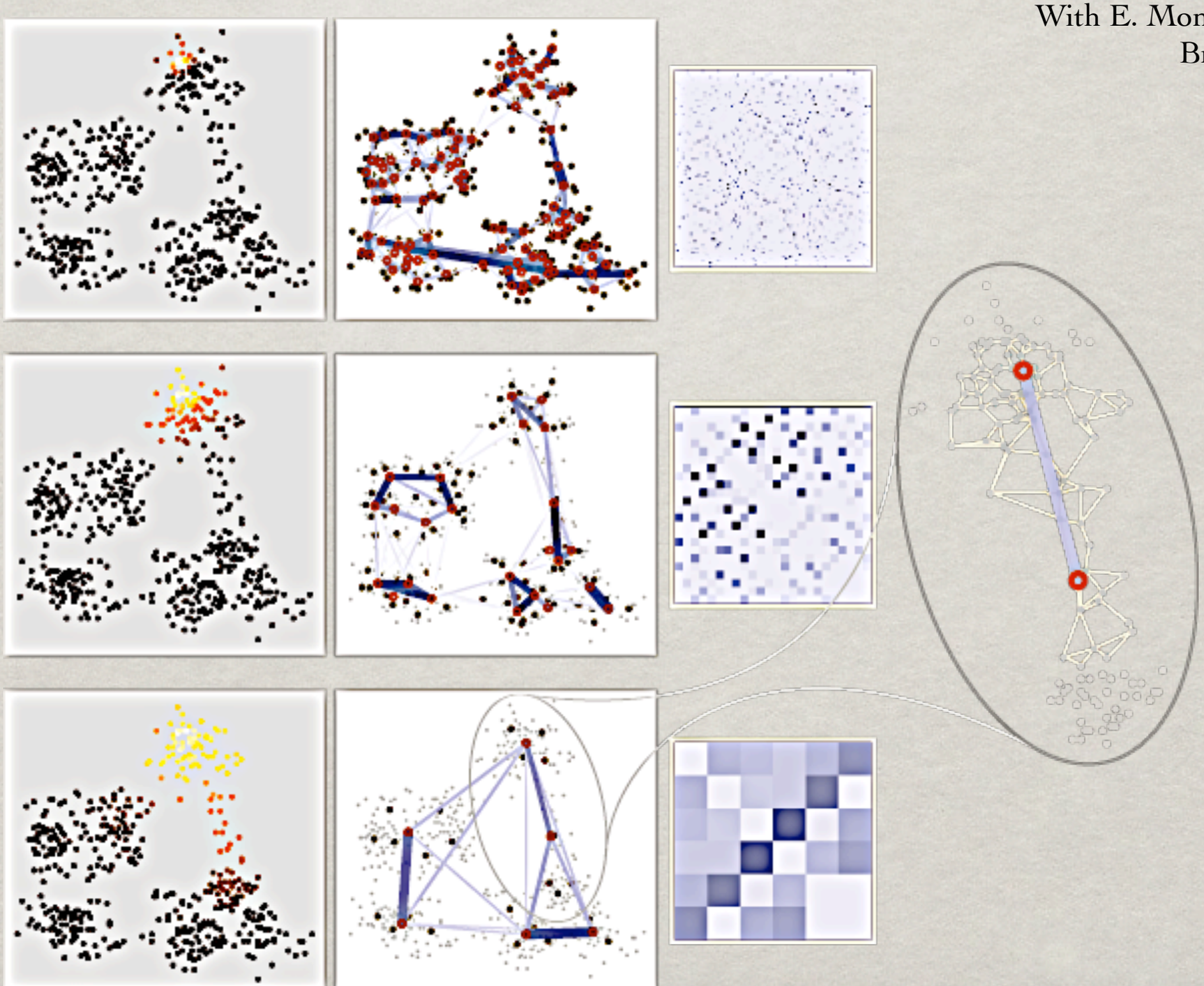
We would like to be able to perform **multiscale analysis of** graphs, and of functions **on** graphs.

Of: produce coarser and coarser graphs, in some sense sketches of the original at different levels of resolution. This could allow a multiscale study of the geometry of graphs.

On: produce coarser and coarser functions on graphs, that allow, as wavelets do in low-dimensional Euclidean spaces, to analyse a function at different scales. We tackle these two questions at once.

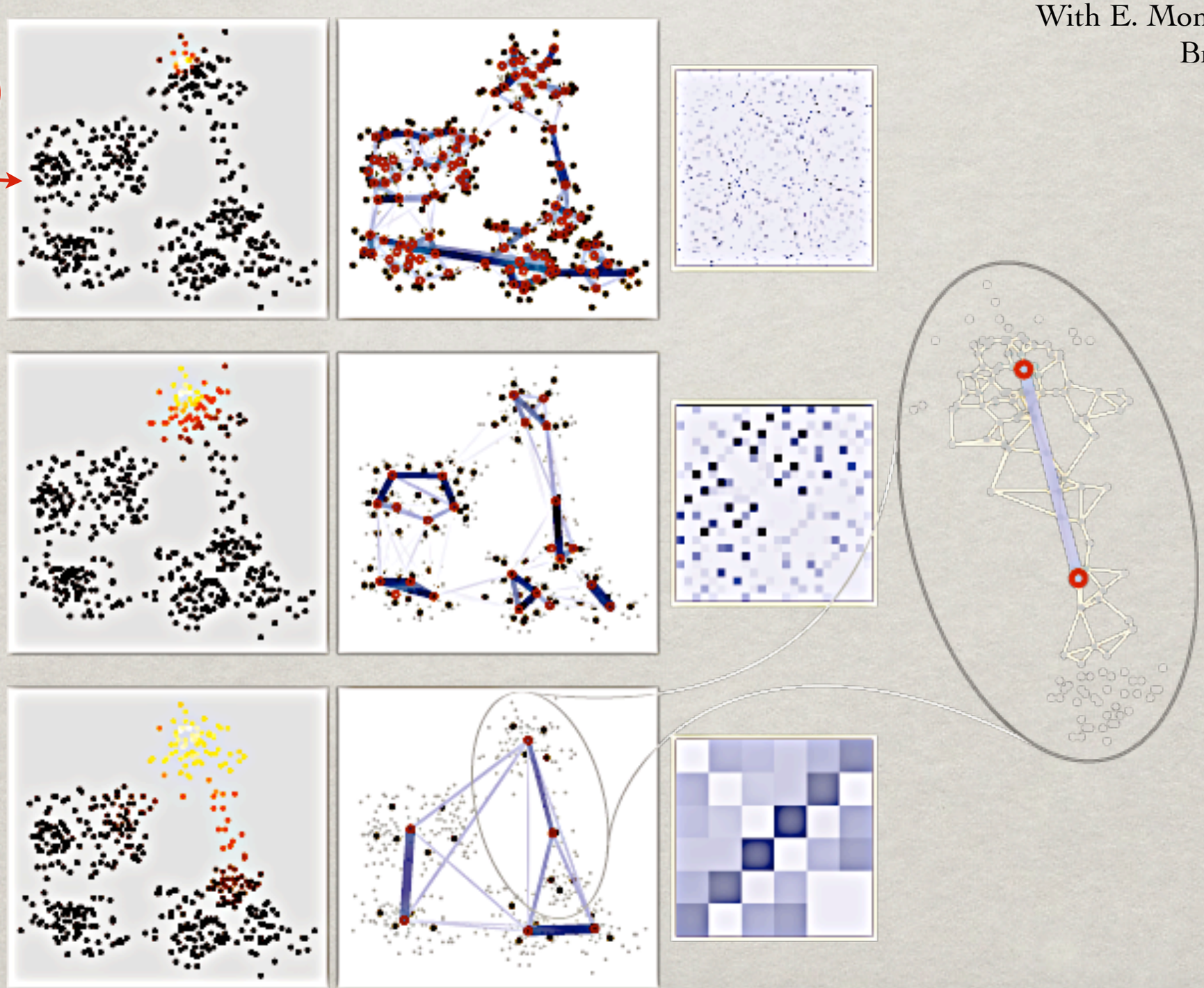
Multiscale graph representations

With E. Monson and R. Brady [C.S.]



Multiscale graph representations

Simple geometric graph from a 2-D point cloud



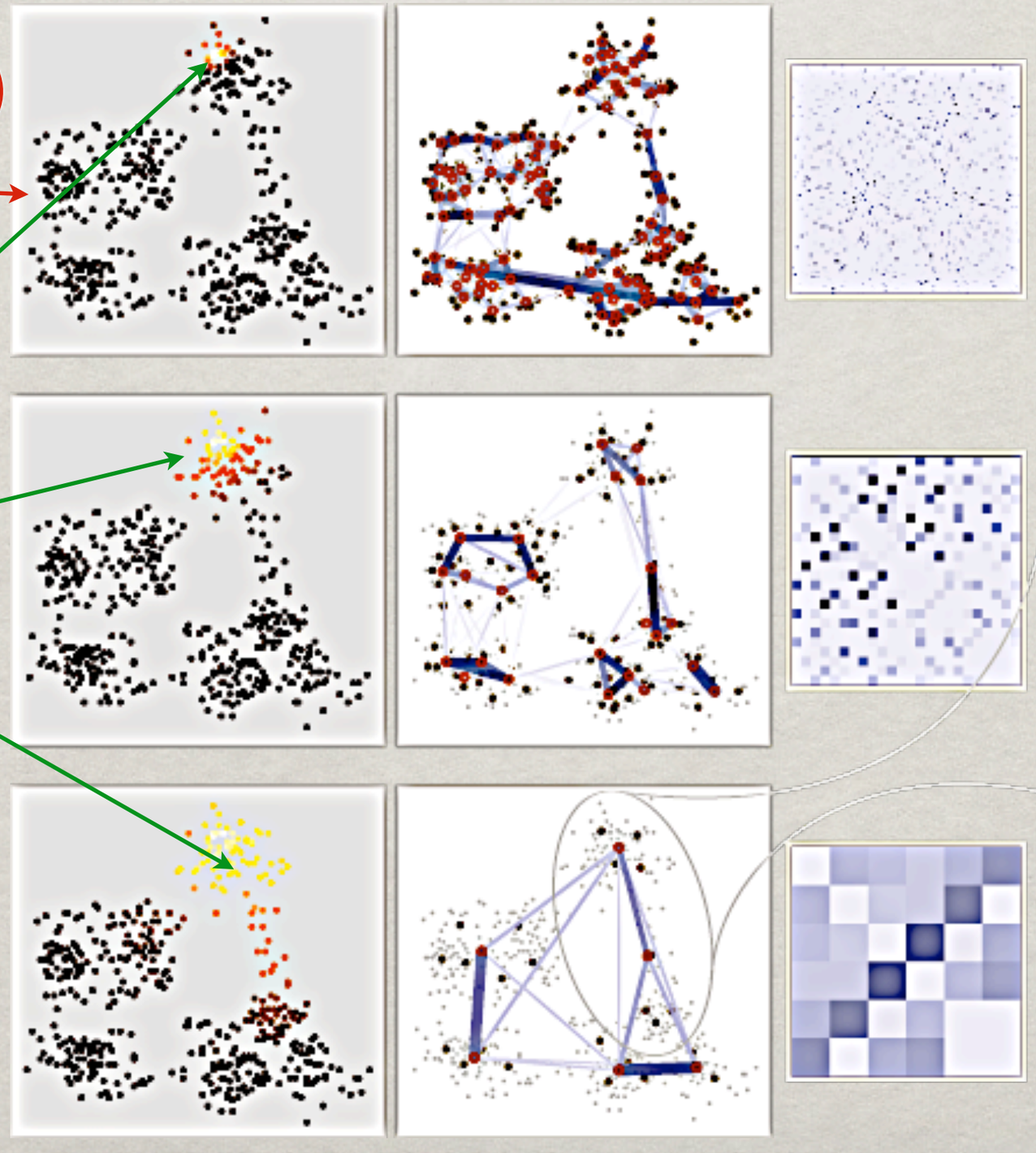
With E. Monson and R. Brady [C.S.]

Multiscale graph representations

With E. Monson and R. Brady [C.S.]

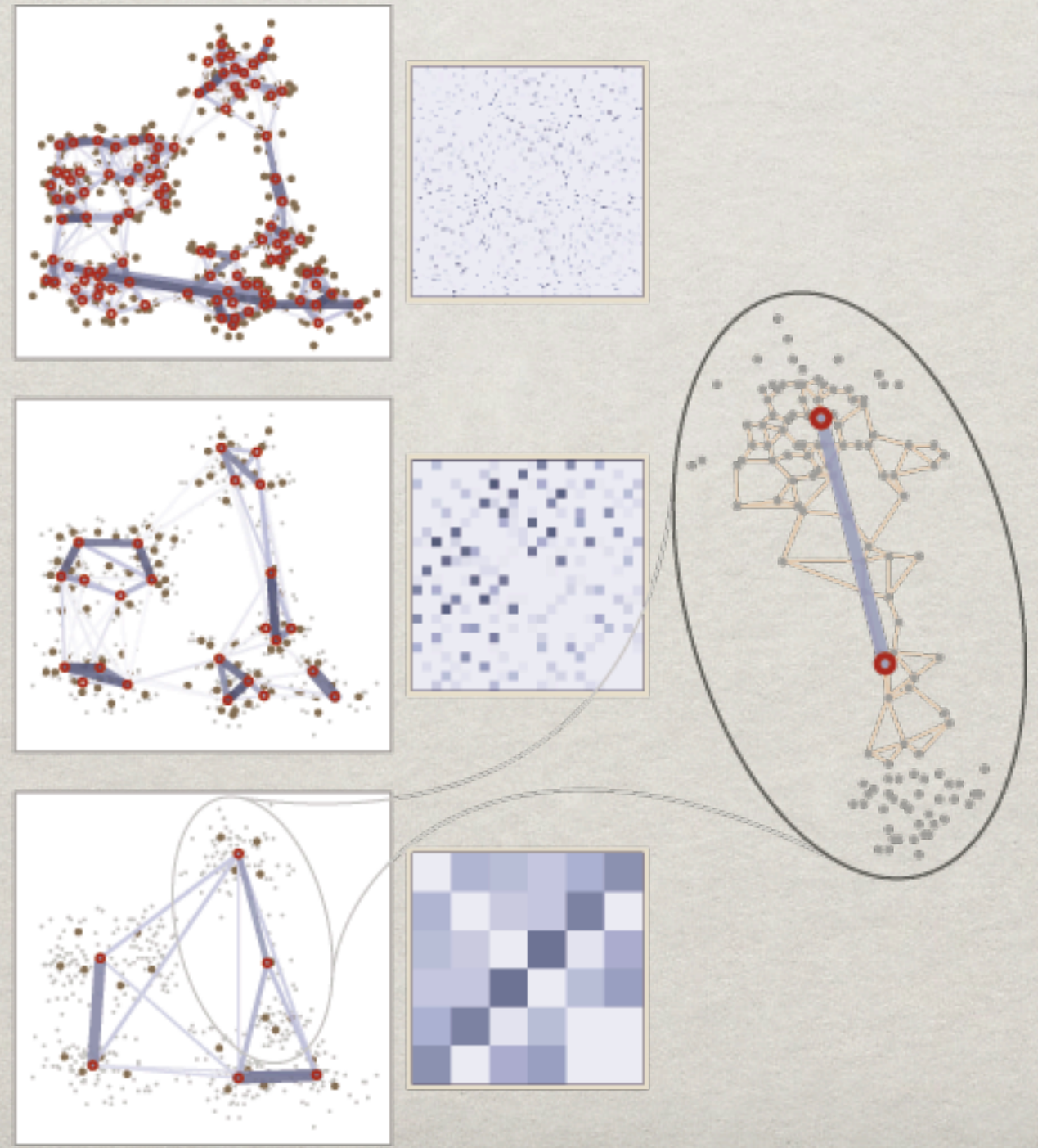
Simple geometric graph from a 2-D point cloud

Functions at multiple scales



Multiscale graph representations

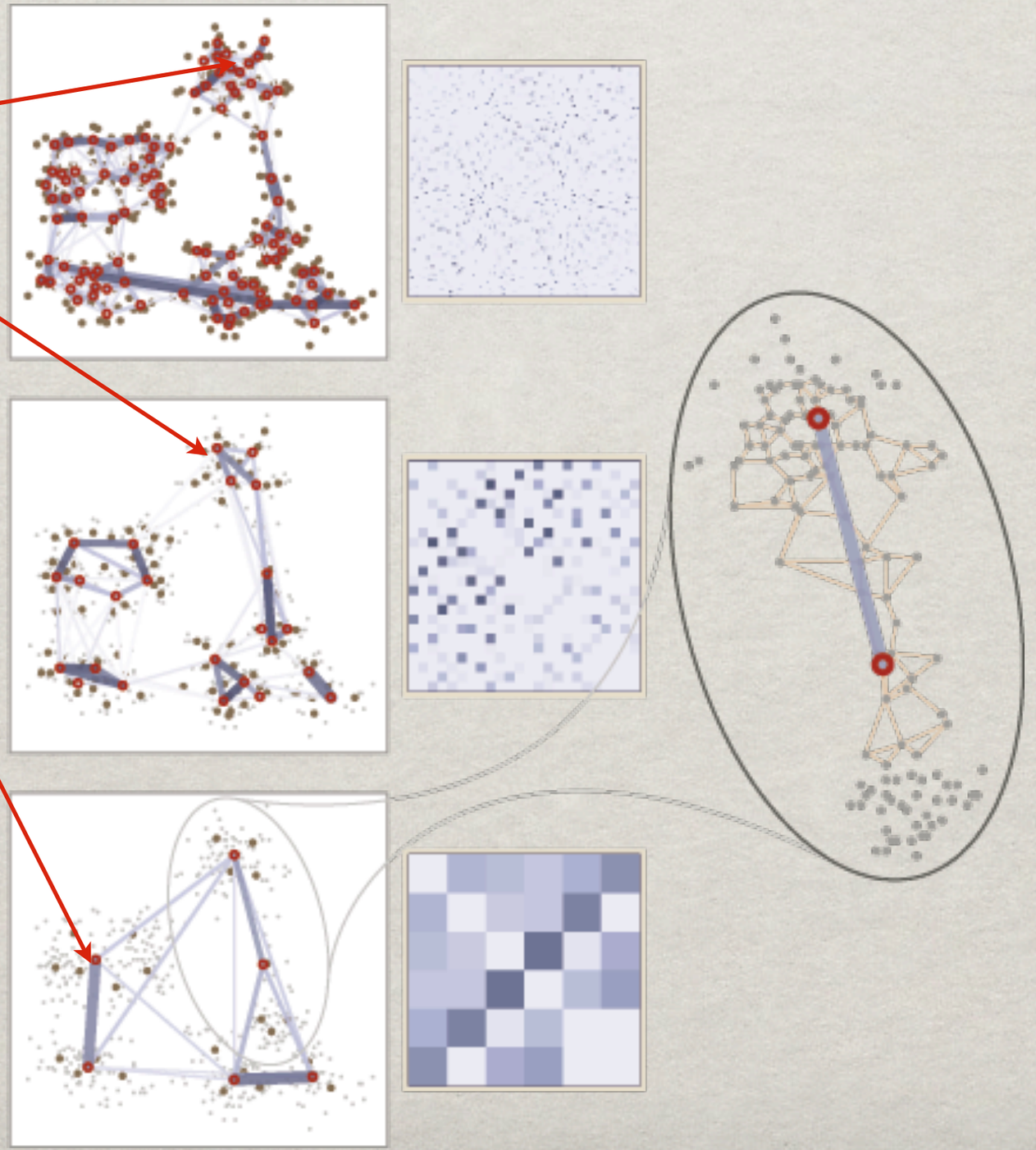
With E. Monson and R. Brady [C.S.]



Multiscale graph representations

With E. Monson and R. Brady [C.S.]

Each function has a center vertex (red dots)

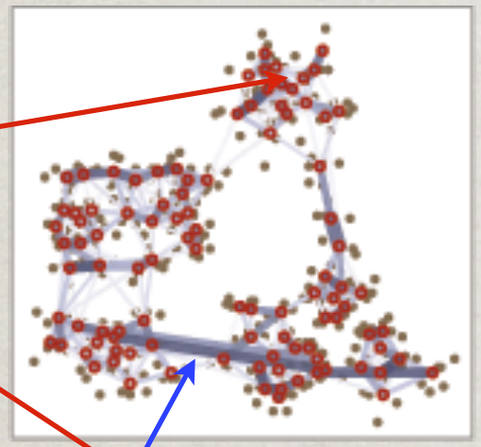


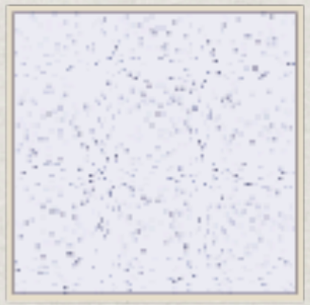
Fine to coarse -Fewer functions
Greater compression - Smaller r.w. matrices

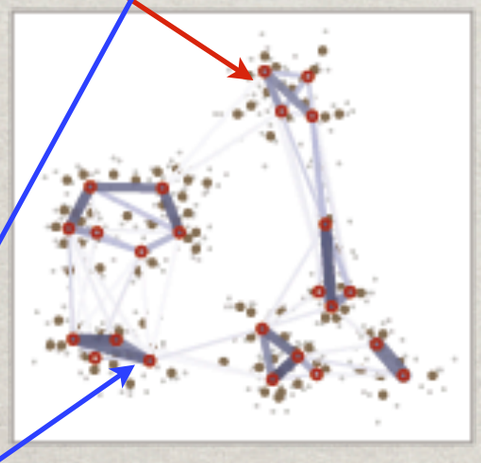
Multiscale graph representations

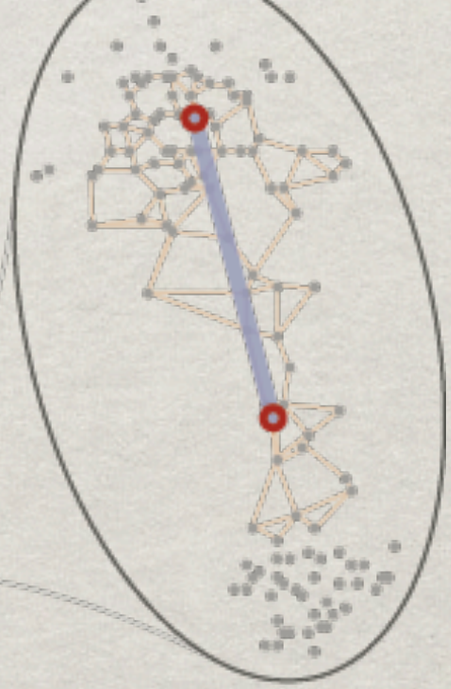
With E. Monson and R. Brady [C.S.]

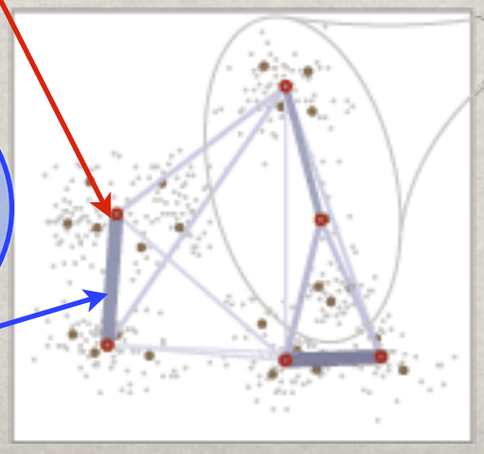
Each function has a center vertex (red dots)











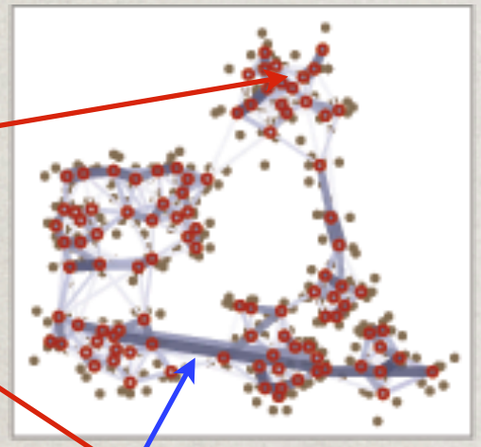
Random walks between functions dictate connectivity and diffusion rates

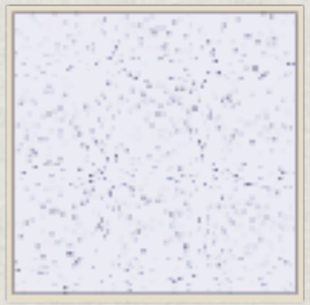
Fine to coarse - Fewer functions
Greater compression - Smaller r.w. matrices

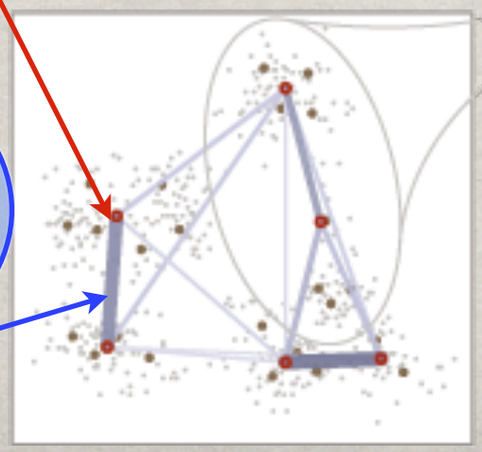
Multiscale graph representations

With E. Monson and R. Brady [C.S.]

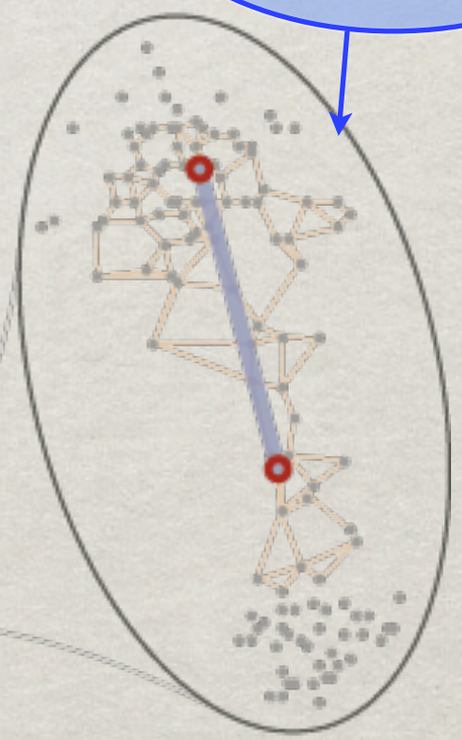
Each function has a center vertex (red dots)







Each edge has a diffusion rate that summarizes that of many finer edges



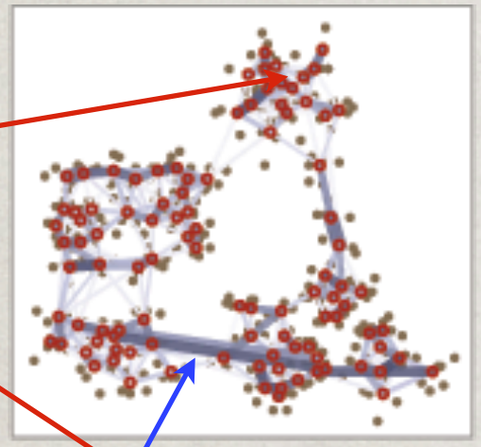
Random walks between functions dictate connectivity and diffusion rates

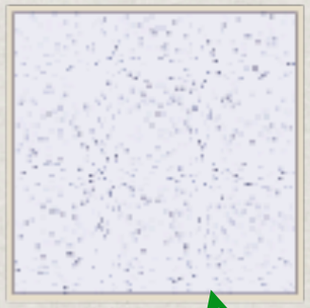
Fine to coarse - Greater compression - Fewer functions - Smaller r.w. matrices

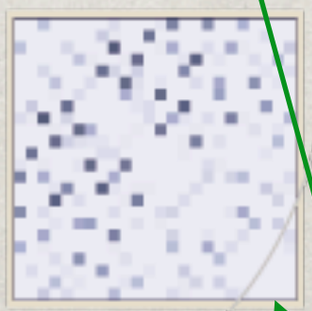
Multiscale graph representations

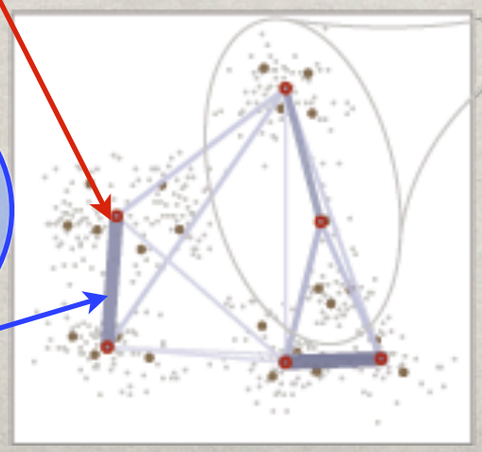
With E. Monson and R. Brady [C.S.]

Each function has a center vertex (red dots)

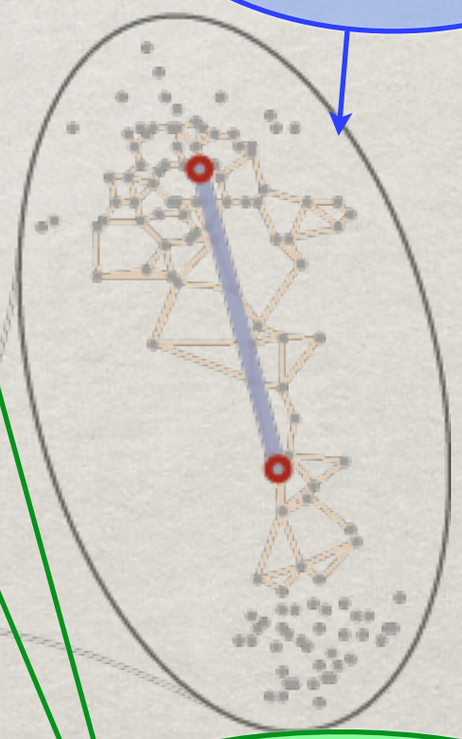








Each edge has a diffusion rate that summarizes that of many finer edges



Random walks between functions dictate connectivity and diffusion rates

Multiscale random walk matrices

Fine to coarse - Greater compression - Smaller r.w. matrices - Fewer functions

Multiscale random walks

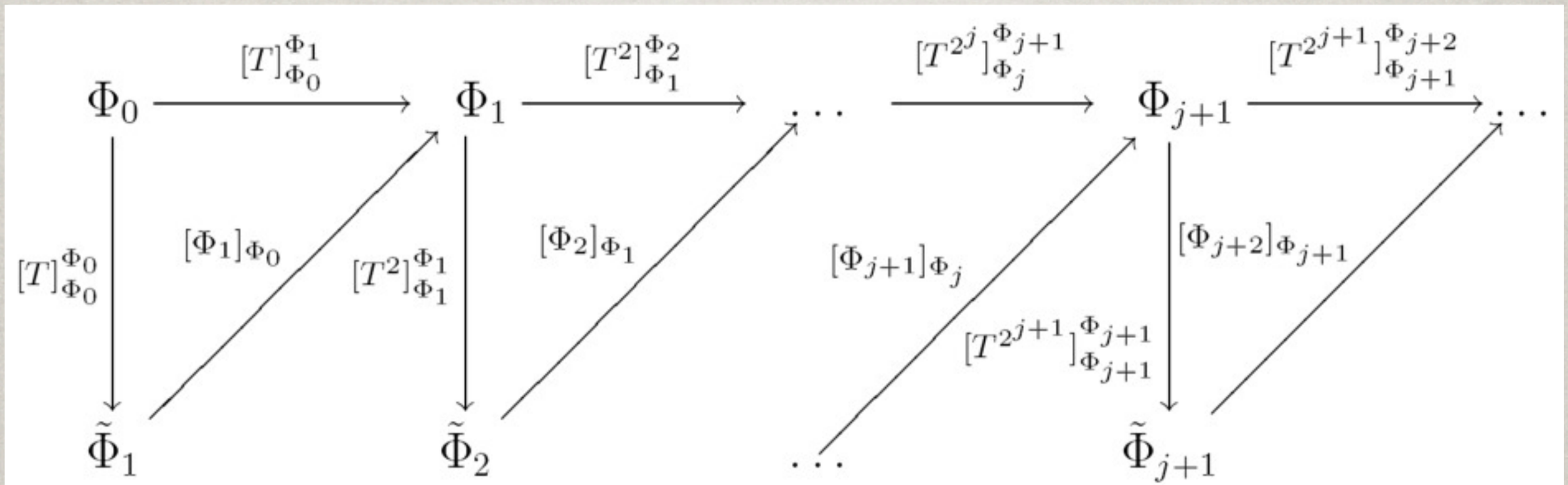
We construct multiscale analyses associated with a diffusion-like process T on a space X, be it a manifold, a graph, or a point cloud. This gives:

(i) A coarsening of X at different "geometric" scales, in a chain $X \to X_1 \to X_2 \to \cdots \to X_j \ldots$;

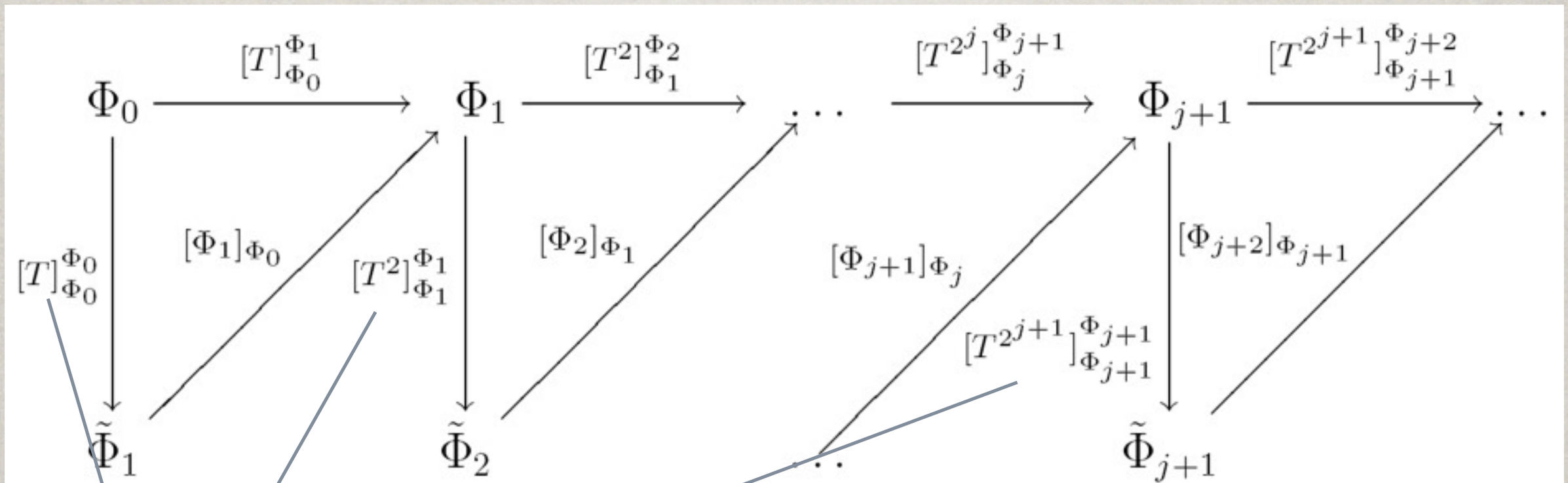
(ii) A coarsening (or compression) of the process T^t at all time scales $t = t_j = 2^j$, $\{T_j = [T^{2^j}]_{\Phi_j}^{\Phi_j}\}_j$, each acting on the corresponding X_j;

(iii) A set of wavelet-like basis functions for analysis of functions (observables) on the manifold/graph/point cloud/set of states of the system.

Scheme for MRA

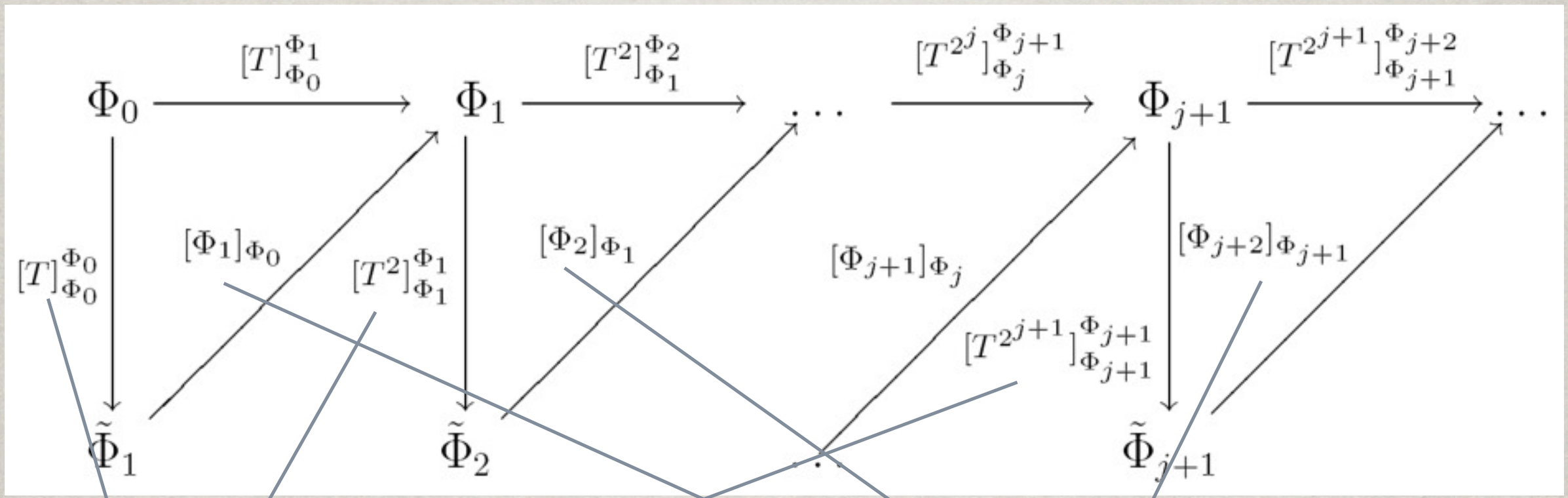


Scheme for MRA



Dilations

Scheme for MRA

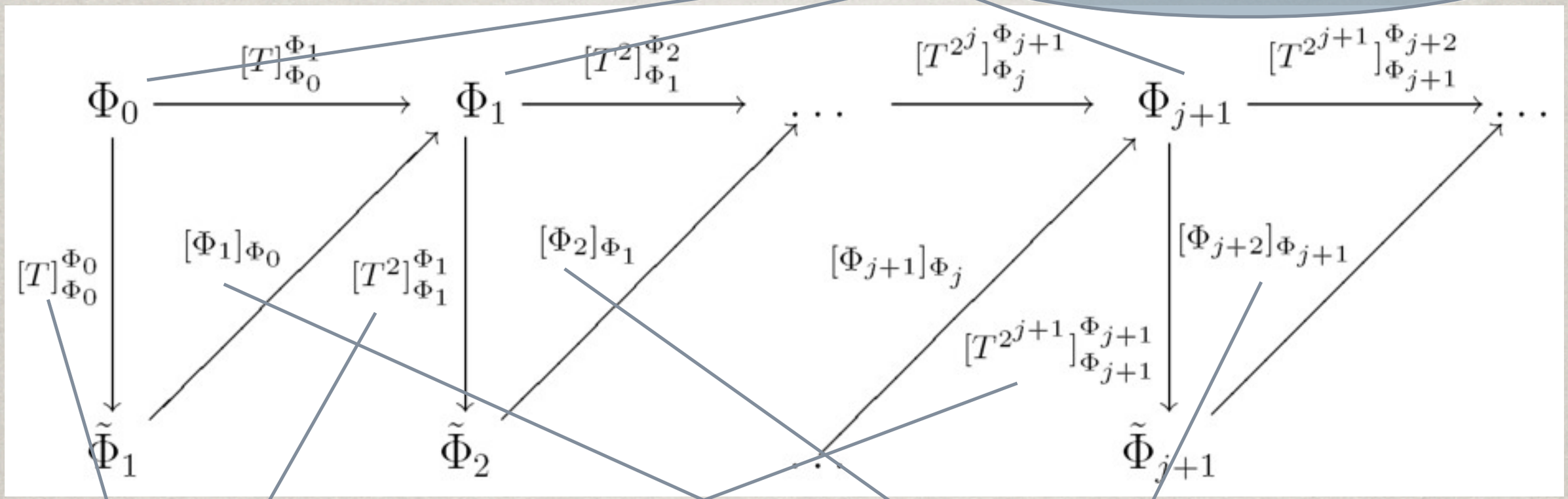


Dilations

Compression

Scheme for MRA

Scaling functions and dilations



Dilations

Compression

Compression step: more details

In order to compress the matrix T we use "rank-revealing QR" decompositions. Fix $\epsilon > 0$.

$$T\Pi = QR = \left(\begin{array}{c|c} Q_{11} & Q_{12} \end{array} \right) \left(\begin{array}{c|c} R_{11} & R_{12} \\ \hline 0 & R_{22} \end{array} \right) \simeq_\epsilon Q_{11} \left(\begin{array}{c|c} R_{11} & R_{12} \end{array} \right)$$

- Q orthogonal, R upper triangular,Π permutation, $||R_{22}||_2 \simeq \epsilon$
- Q are the scaling functions $[\Phi_1]_{\Phi_0}$, $[R_{11}|R_{12}]$ is $[T]_{\Phi_0}^{\Phi_1}$, the compressed operator from fine to coarse scale.
- The number of columns N_1 of Q_{11} (and of R_{11}) determines the dimension of the next coarse scale.
- The first N_1 columns of Π select N_1 representative vertices on the graph.

Consistence of multiscale r.w.'s

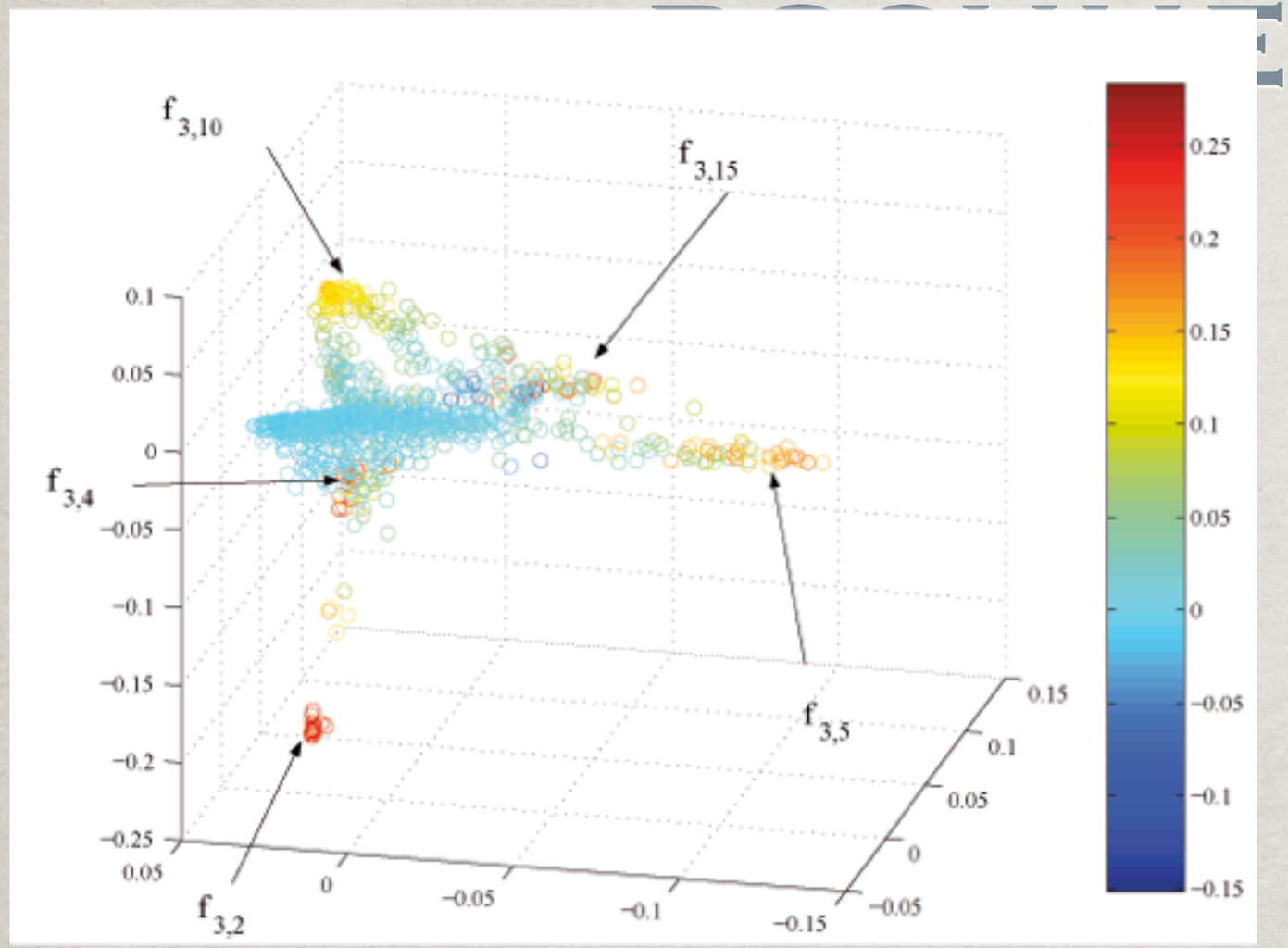
Let $G_j = \Phi_j$ be the graph whose vertices are the scaling functions at scale j. $T_j \approxeq [T^{2^{j+1}}]_{\Phi_j}^{\Phi_j}$ is a "random walk" (symmetrized) on G_j, a compressed version of T^{2^j} restricted to V_j. In fact:

$$T^{2^{j+1}} f \approxeq P_{V_j}^* T_j P_{V_j} f$$

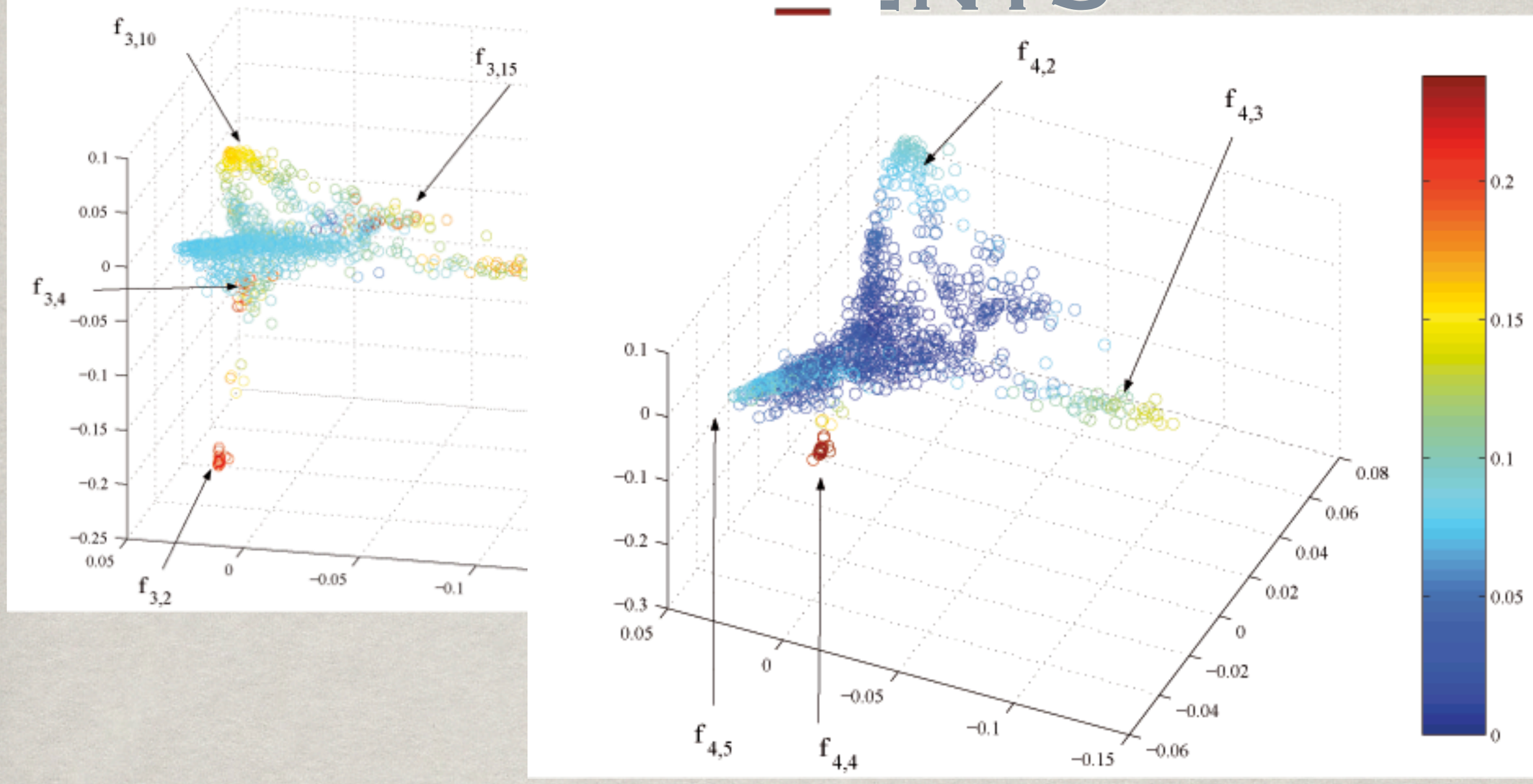
with $\approxeq$ meaning ϵ-close in $L^2(G)$.
We think of (G_j, T_j) as a coarse version of G, constructed in such way that the random walk on G_j is the random talk on G at time 2^j, compressed.

Example 2: text documents

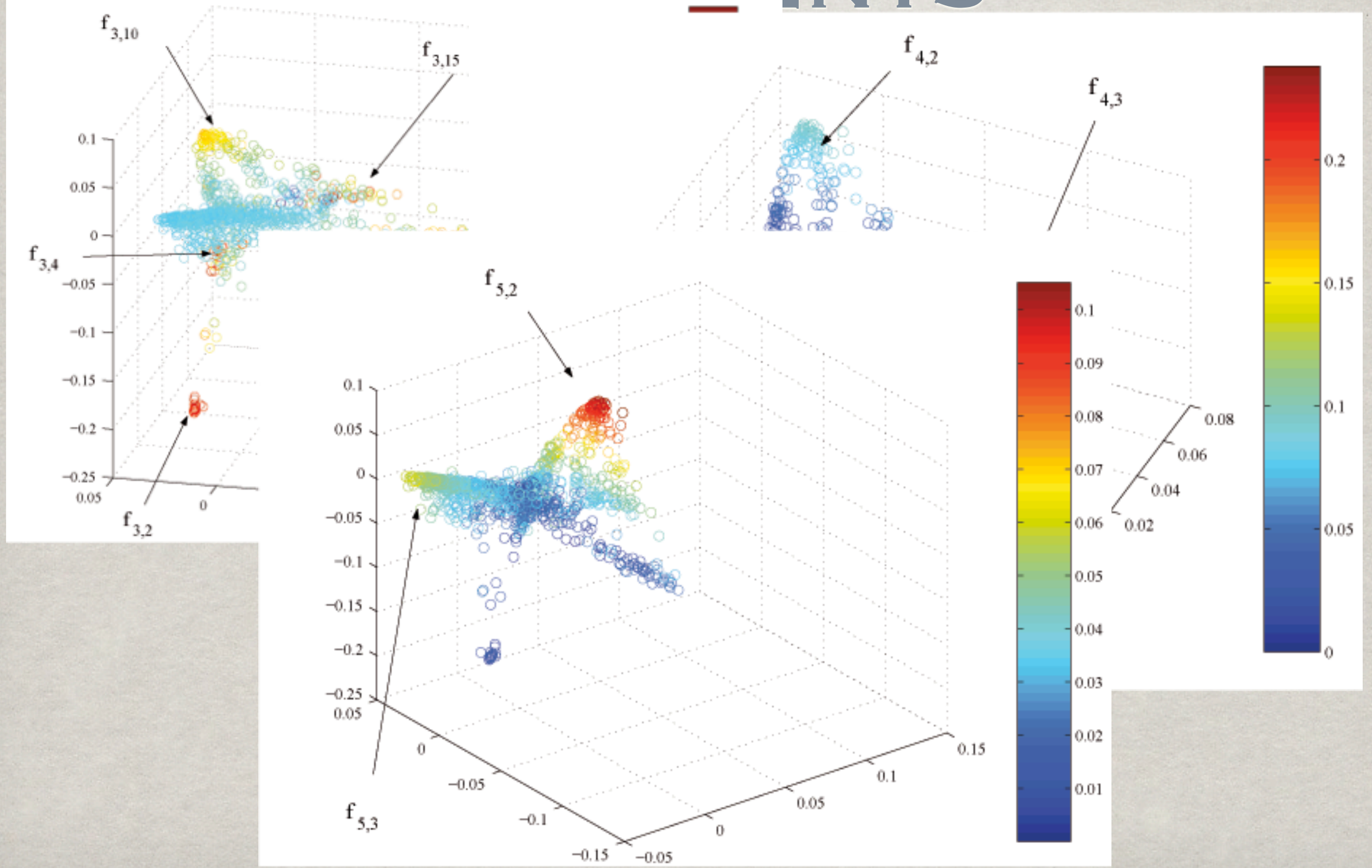
Example 2: text documents



Example 2: text documents



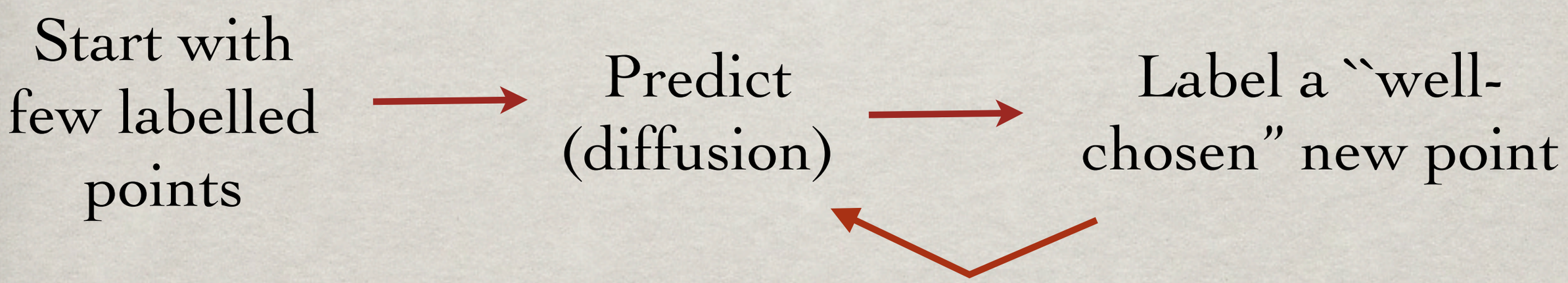
Example 2: text documents



Active Learning

With E. Monson and R. Brady [C.S.]

We observe the full data set (e.g. a body of text documents), and want to learn a categorization of the data (e.g. topics of the text documents). We pay a price for every label we obtain from an expert.
We would like to find points s.t. if we get their correct label we maximize the gain in prediction accuracy.
Natural candidates are the multiscale diffusion centers. We compare them to using random points and to using points of high degree.

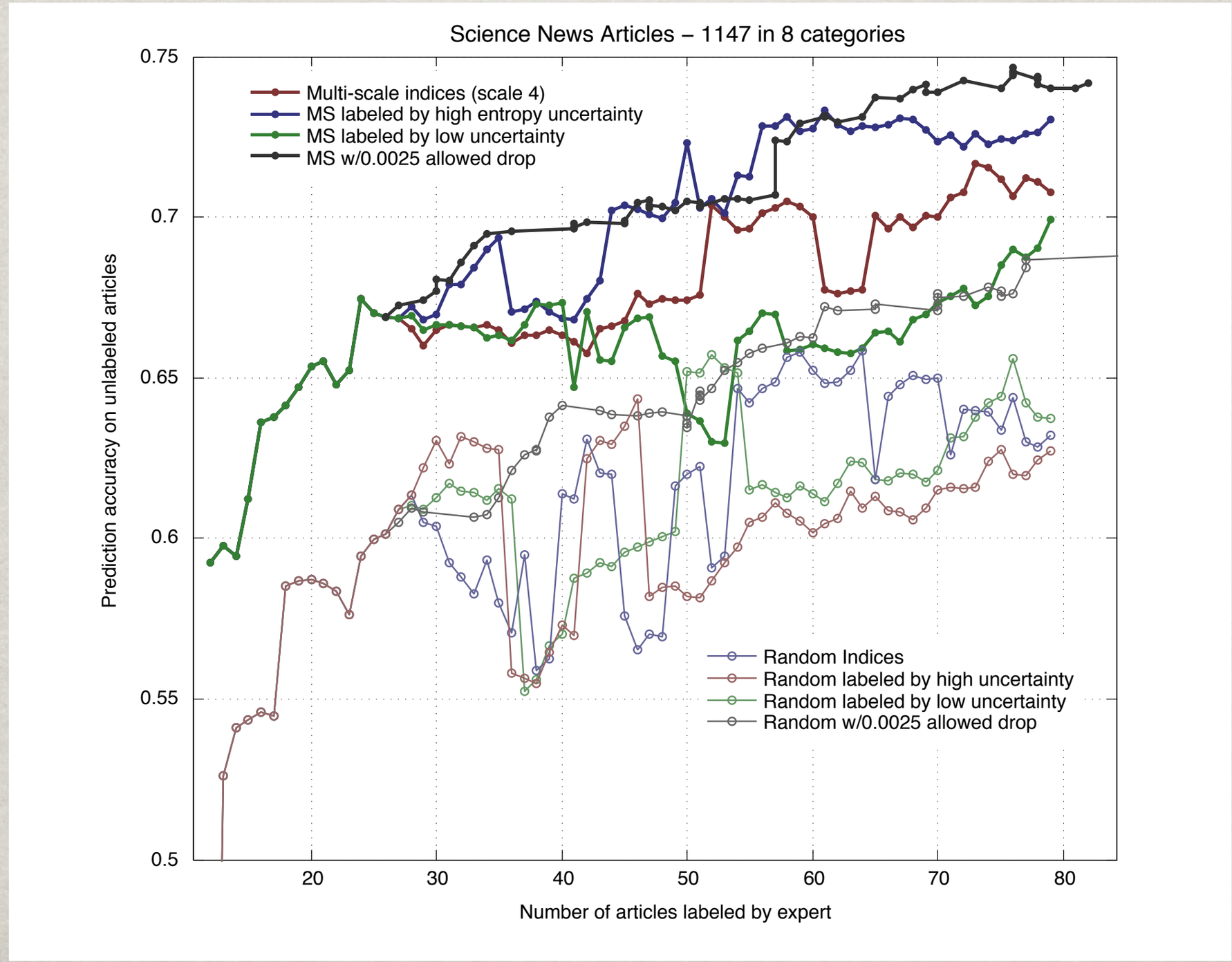


``Well-chosen": well-spread, and with highly uncertain prediction

Active Learning

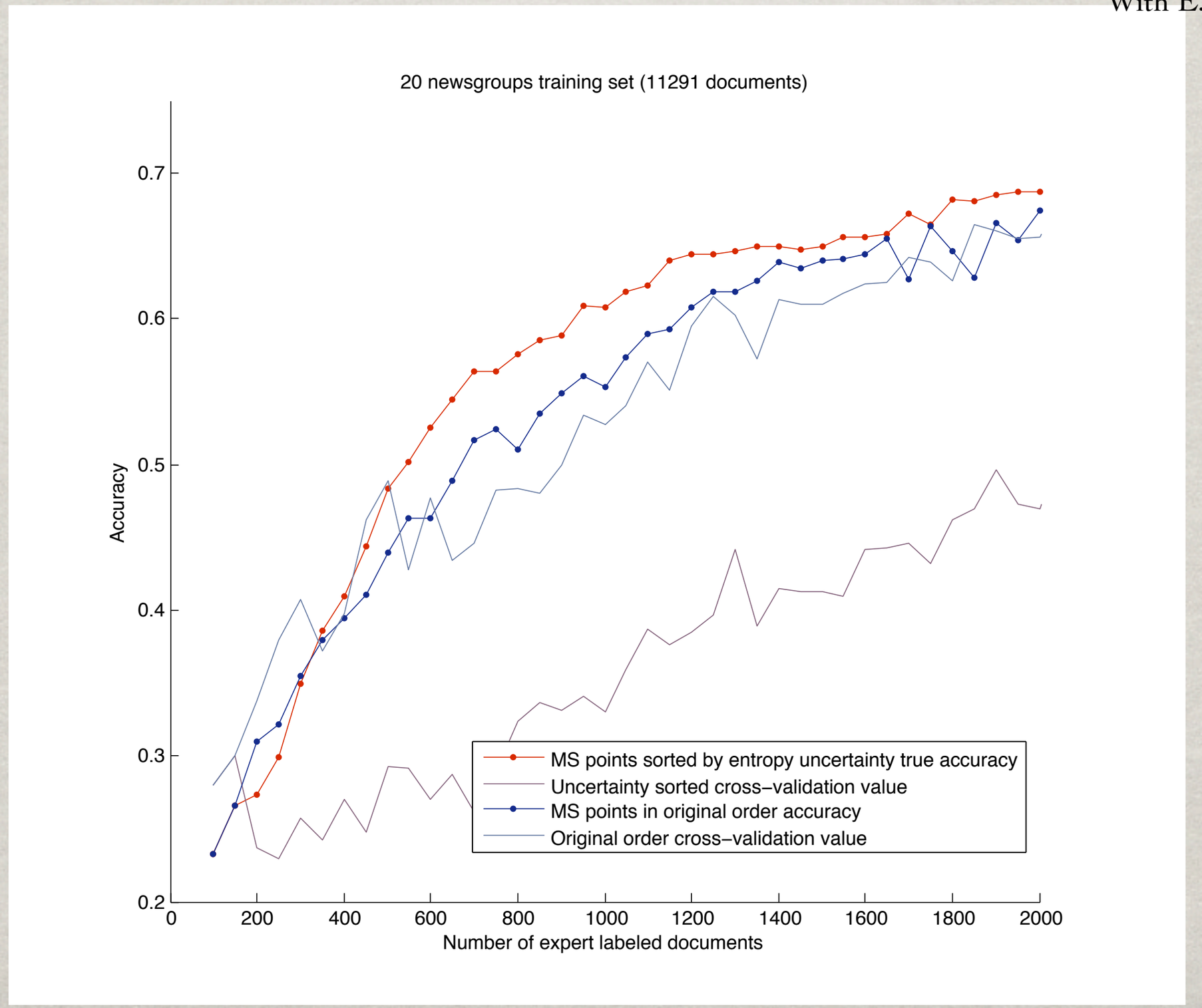
1147 Science News articles, 8 categories

With E. Monson and R. Brady [C.S.]



Active Learning

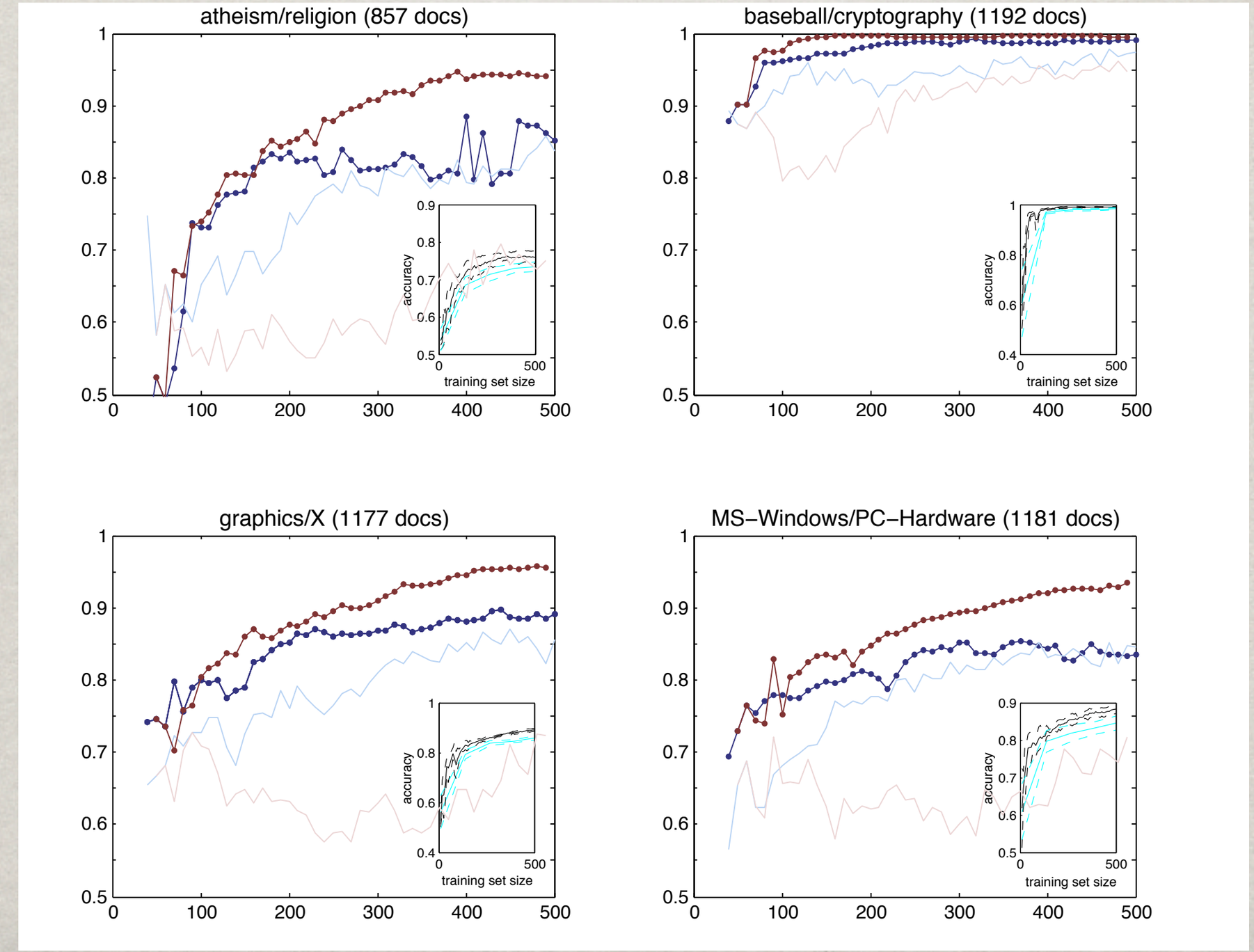
With E. Monson and R. Brady [C.S.]



Active Learning

11291 Newsgroup Data, 20 categories

With E. Monson and R. Brady [C.S.]



Insets from: Schohn, G. Cohn, D. (2000) Less is more: Active learning with support vector machines, Proceedings of the Seventeenth International Conference on Machine Learning, (ICML-2000), 839-846, Stanford, CA.

Multiscale inference

With J. Guinney [Comp.Bio.], S. Mukherjee [Stat.], P. Febbo [Med.]

A data matrix X, $N \times D$ may be thought as representing data on a δ-function basis (N single samples in terms of D single coordinates).

When we compute the SVD $X = U\Sigma V^T$ we are computing global coordinates in sample space (the columns of U) and in coordinate space (the columns of V^T). It is a sort of linear Fourier analysis.

With this multiscale decomposition we can interpolate in-between at all scales and obtain useful data representations.

The random walk methods do that nonlinearly, by following the geometry of the data. Equivalent to working in a higher-dimensional feature space inferred from the data.

Multiscale inference

With J. Guinney [Comp.Bio.], S. Mukherjee [Stat.], P. Febbo [Med.]

X is $N \times D$, N documents in $\mathbb{R}^D$, compute multiscale dictionary Φ $(D \times M)$ on the D words. If f maps documents to their topic, write $f = X\Phi\beta + \eta$ and find β by

$$\text{argmin}_\beta ||f - X\Phi\beta||_2^2 + \lambda ||\{2^{-j\gamma}\beta_{j,k}\}||_1 ,$$

which is a form of sparse regression. (λ, γ) are determined by cross-validation.

Fitting term, with linear function of multiscale features

Sparsity in terms of multiscale dictionary

We obtain very sparse solutions on various data sets, with corresponding basis elements having different scales.

Example: text documents

X is $N \times D$, N documents in $\mathbb{R}^D$, compute multiscale dictionary Φ ($D \times M$) on the D words. If f maps documents to their topic, write $f = X\Phi\beta + \eta$ and find β by

$$\text{argmin}_\beta ||f - X\Phi\beta||_2^2 + \lambda ||\{2^{-j\gamma}\beta_{j,k}\}||_1 \,,$$

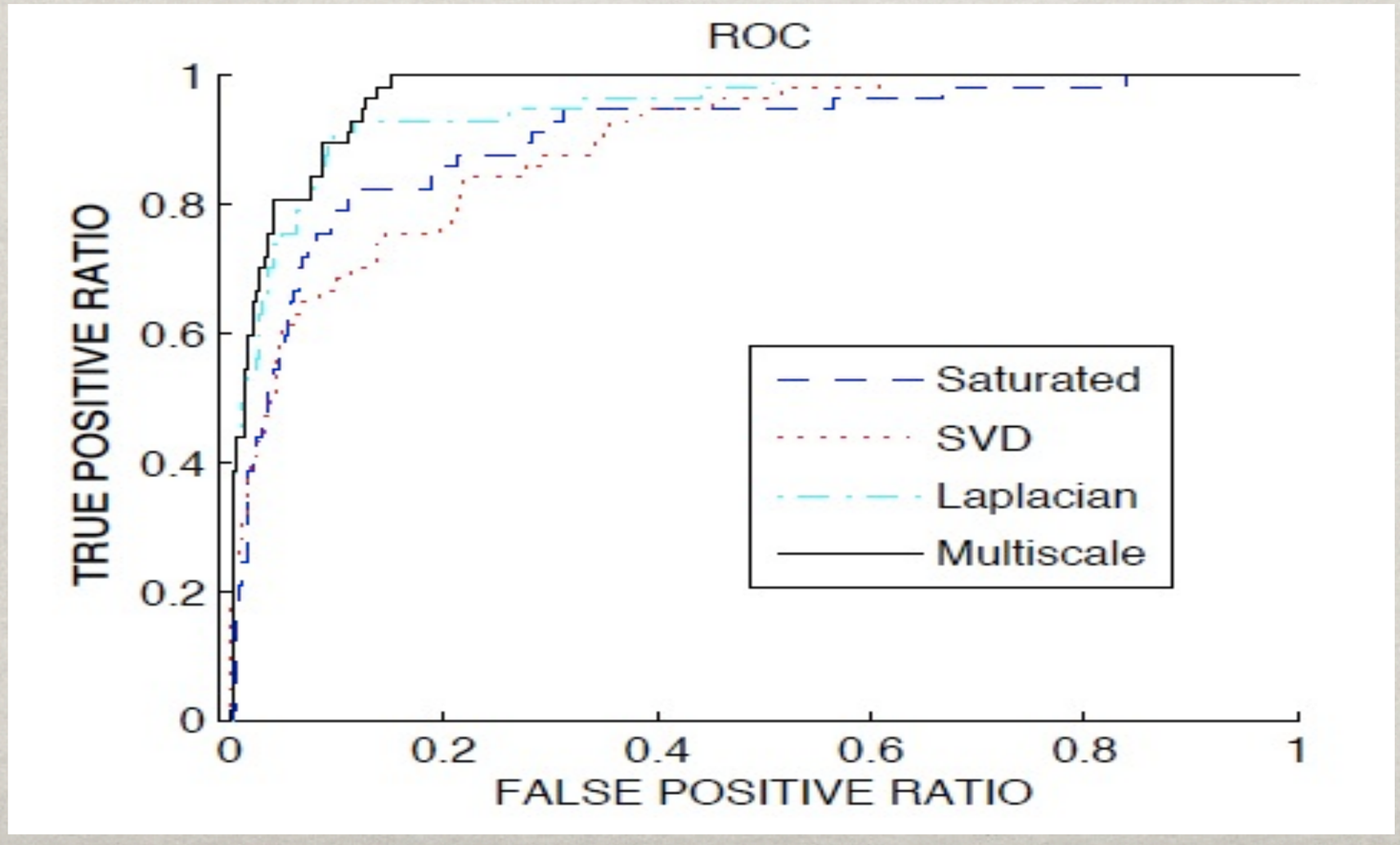
which is a form of sparse regression. (λ, γ) are determined by cross-validation.

Example: text documents

X is $N \times D$, N documents in $\mathbb{R}^D$, compute multiscale dictionary Φ ($D \times M$) on the D words. If f maps documents to their topic, write $f = X\Phi\beta + \eta$ and find β by

$$\text{argmin}_\beta ||f - X\Phi\beta||_2^2 + \lambda ||\{2^{-j\gamma}\beta_{j,k}\}||_1 \,,$$

which is a form of sparse regression. (λ, γ) are determined by cross-validation.



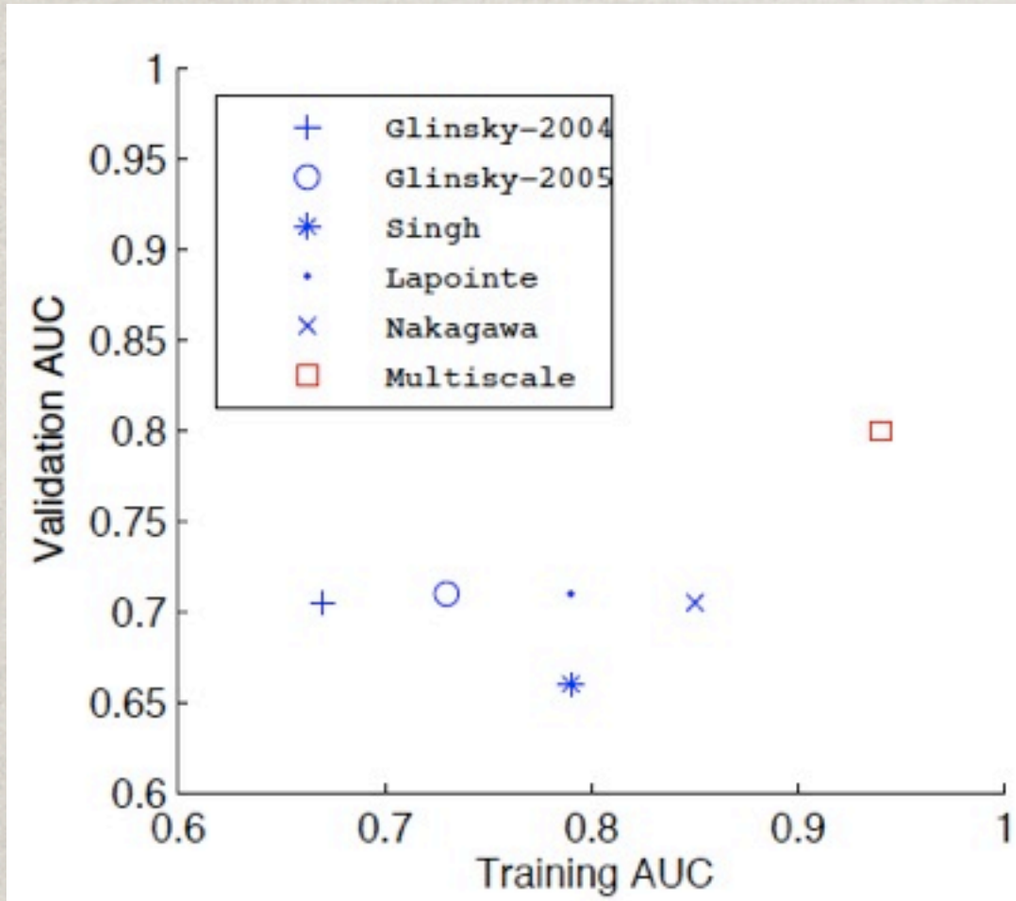
Example: gene arrays

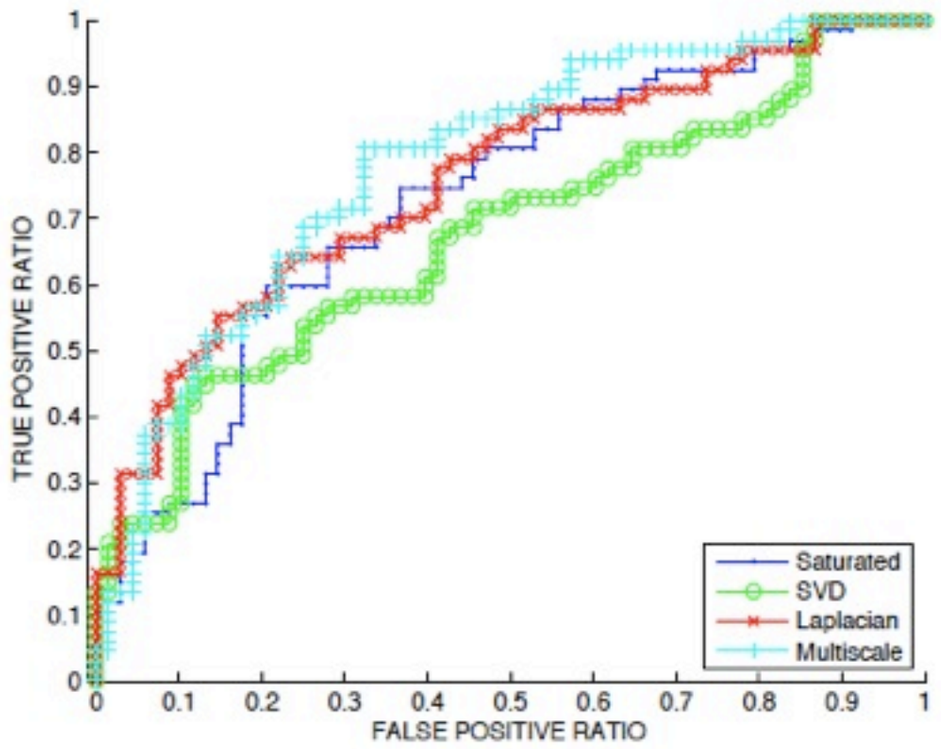
X is $N \times D$, N patients with D genes (here $N \sim 400$ and $D \sim 1000$).

Source of data: Nakagawa T, Kollmeyer T, Morlan B, Anderson, S, Bergstralh E, et al, (2008)
A tissue biomarker panel predicting systemic progression after PSA recurrence post-definitive prostate cancer therapy, Plos One 3:e2318.

Example: gene arrays

X is $N \times D$, N patients with D genes (here $N \sim 400$ and $D \sim 1000$).

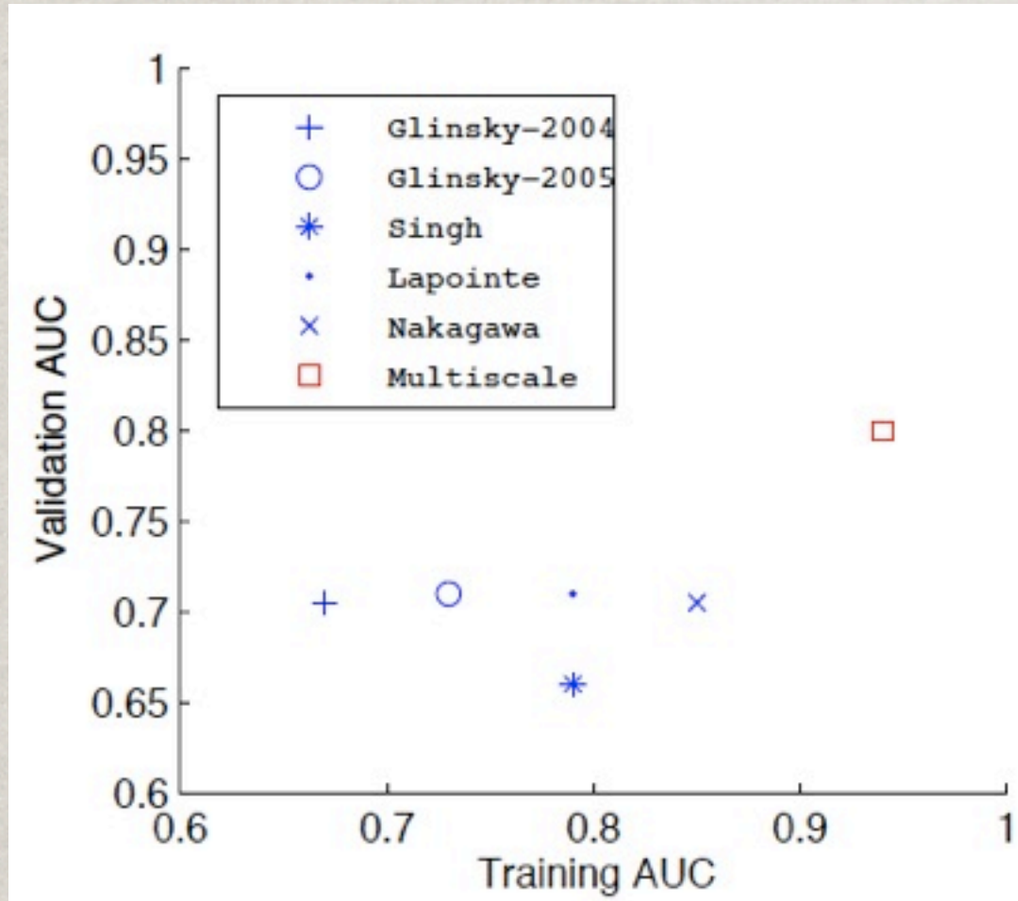


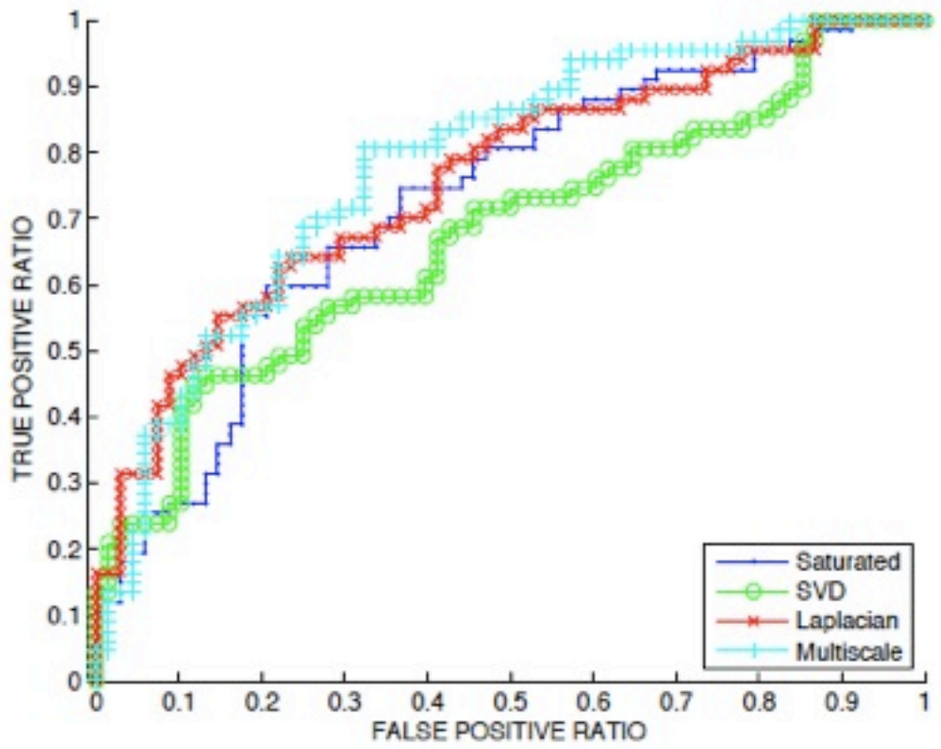


Source of data: Nakagawa T, Kollmeyer T, Morlan B, Anderson, S, Bergstralh E, et al, (2008)
A tissue biomarker panel predicting systemic progression after PSA recurrence post-definitive prostate cancer therapy, Plos One 3:e2318.

Example: gene arrays

X is $N \times D$, N patients with D genes (here $N \sim 400$ and $D \sim 1000$).





Added advantage: the multiscale genes we construct are much interpretable than eigengenes, several of them match important pathways, and moreover both small scale and large scale genelets seem relevant.

Source of data: Nakagawa T, Kollmeyer T, Morlan B, Anderson, S, Bergstralh E, et al, (2008) A tissue biomarker panel predicting systemic progression after PSA recurrence post-definitive prostate cancer therapy, Plos One 3:e2318.

Heat and Eigenfunction maps

With P.W. Jones and R. Schul [Math.]

We have results that show that one may use heat kernels or Laplacian eigenfunctions to obtain bi-Lipschitz maps from large portion of a manifold $\mathcal{M}$ of dimension k to Euclidean space $\mathbb{R}^k$. These manifold learning algorithms are quite different from existing ones. Maps are in the form

- $x \mapsto (K_t(x, x_i))_{i=1}^k$ where $K_t(\cdot, x_i)$ is the heat kernel on $\mathcal{M}$ centered at $x_i \in \mathcal{M}$ at time t. The sources x_i are well-chosen depending on the large region being mapped.

- $x \mapsto (\varphi_{j_i}(x))_{i=1}^k$ where the φ_j's are eigenfunctions of the Laplacian Δ on $\mathcal{M}$, and the indices j_i are well-chosen, depending on the large region of $\mathcal{M}$ being mapped.

These algorithms, as most others, require k as in input.

Heat and Eigenfunction maps

With P.W. Jones and R. Schul [Math.]

These results require two parameters:

(a) The radius R of the largest ball that admits a $(1+\epsilon)$-biLipschitz embedding.

(b) The intrinsic dimensionality k.

(a) seems hard, but in fact trivial by greedy multiscale algorithm, that applies the Theorem for $R = R_j := (1+\delta)^j$, for $j = 0, \ldots, J$. If Theorem yields the desired map at scale R_j, increase j, otherwise stop. The Theorem guarantees that we stop at the optimal (oracle) scale (up to a factor $(1+\delta)$).

(b) is a classical problem. Existing algorithms do not seem to perform well and have weak guarantees. Two approaches: use the Theorem (with different k's), or use Multiscale SVD. These two approaches are essentially the same.

Intrinsic dimensionality

With P.W. Jones and R. Schul [Math.]

Model: data $\{x_i\}_{i=1}^n$ is sampled from a manifold $\mathcal{M}$ of dimension k, embedded in $\mathbb{R}^D$, with $k \ll D$. We receive $\{x_i + \eta_i\}_{i=1}^n$, where $\eta_i \sim_{\text{i.i.d}} \eta$ is D-dimensional noise (e.g. Gaussian).
Objective: estimate k. Motivations:

- Basic measure of complexity of the data
- Settle claims about low-dimensional structures in data
- Needed by many algorithms that seek to parametrize the data
- Equivalent to number of: latent variables in a linear model, degrees of freedom in a dynamical system; useful for clustering the data by local dimensionality, finding compressed representations of the data, building dictionaries for representing and modeling the data, etc...
- Much work has been done, but it is quite unsatisfactory (more on this later...)

Rough overview of existing techniques

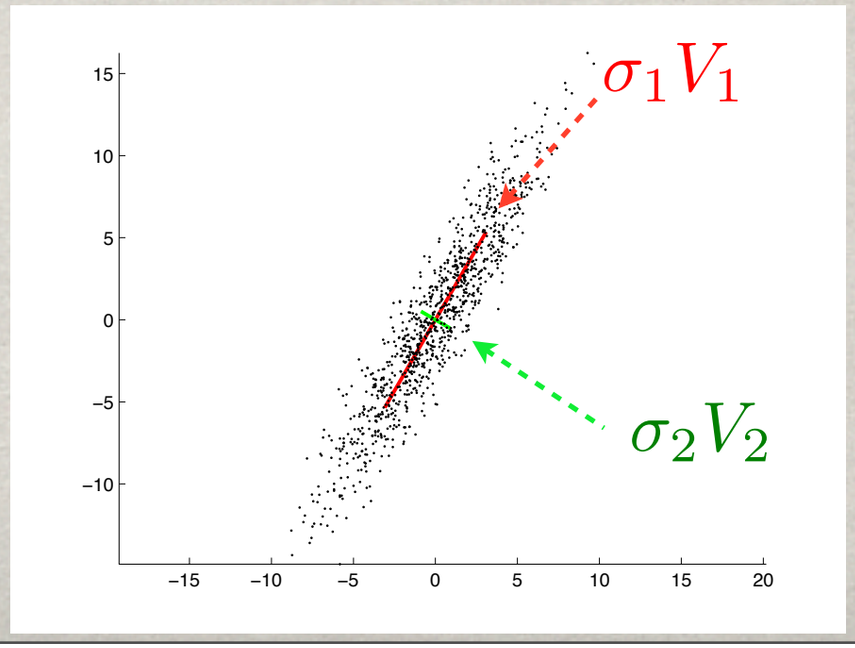
With A. Little [Math.] and L. Rosasco [C.S.]

Two standard methods:

Volume-based: on a k-dimensional set, $|B_r(z) \cap \mathcal{M}| \sim r^k$. Compute $\log |B_r(z)|$ for several values of r and fit a line. Empirical version needs $n \sim 2^k$. Many papers with variations on this theme, including refinements in the last few years. For example one may try to construct $B_r(z)$ *on* the manifold.

Principal Component Analysis: if X_n is the $n \times D$ matrix with the samples, write the Singular Value Decomposition $X = U\Sigma V^T$, where U, V are orthogonal and Σ diagonal with elements (singular values) $\sigma_1 \geq \sigma_2 \geq \dots$. Alternatively, let $\mathrm{cov}(X_n) = \frac{1}{n} X_n^T X_n = \frac{1}{n} V \Sigma^2 V^T$. For n points $x_1, \dots, x_n$, among all i-dimensional planes, the plane π_i spanned by the top i principal vectors V_i minimizes

$$\sum_{l=1}^{n} ||x_l - \pi_i(x_l)||^2 .$$



Covariance estimation results:
exactly k non-zero σ_i's, w.h.p.,
as soon as $n \gtrsim k \log k$.

Sketch of Results

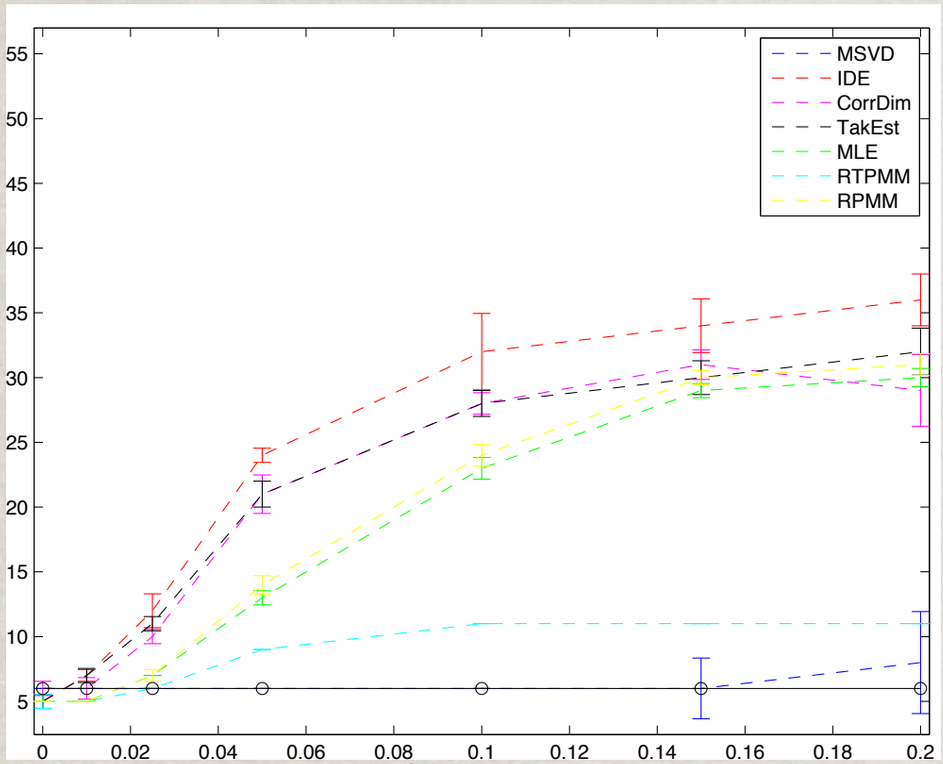
With A. Little [Math.]
and L. Rosasco [C.S.]

We obtain general results which imply, as particular cases, the following: if certain geometric conditions and bounds on the noise hold, the algorithm succeeds

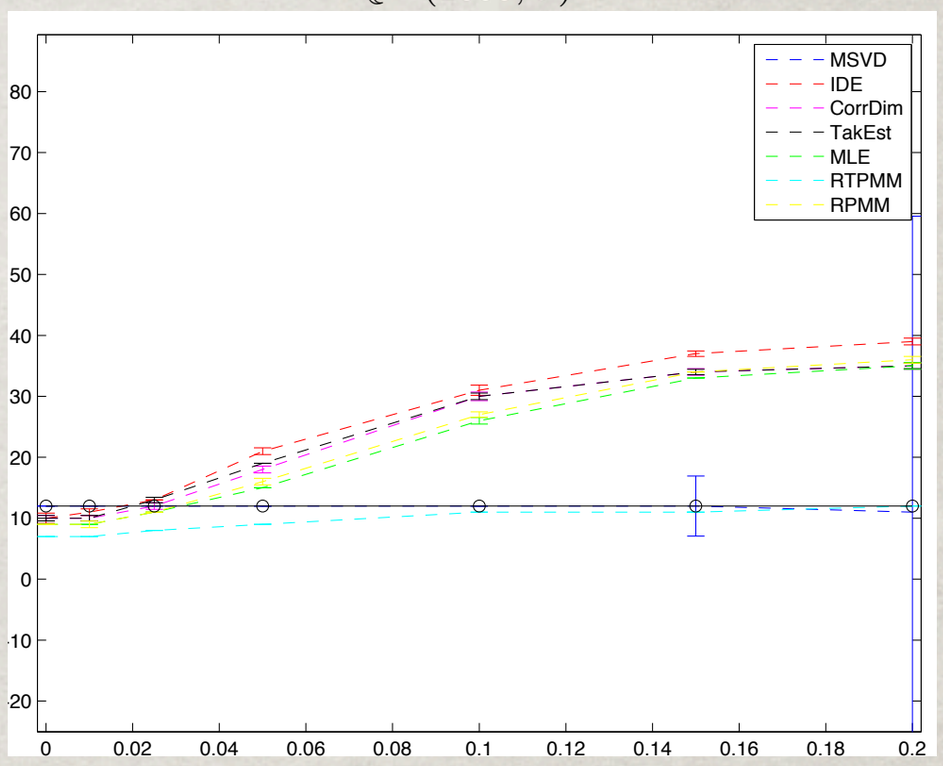
- *Consistency*: as the number of samples $n \to +\infty$
- *Scaling limit*: as $n, D \to +\infty$ with $\frac{n}{D} \to \gamma$
- *(Ambient dimension)-free limit*: for fixed n, for $D \to +\infty$, and $\sigma\sqrt{D} \sim 1$ (i.e. $\mathbb{E}[||\eta||] = O(1)$ independently of D)
- *(Intrinsic dimension)-free*: $n, k \to +\infty$, $\frac{n}{k \log k} \to \gamma$
- *Dimension-free*: a combination of the last two.

Comparisons: unit cube

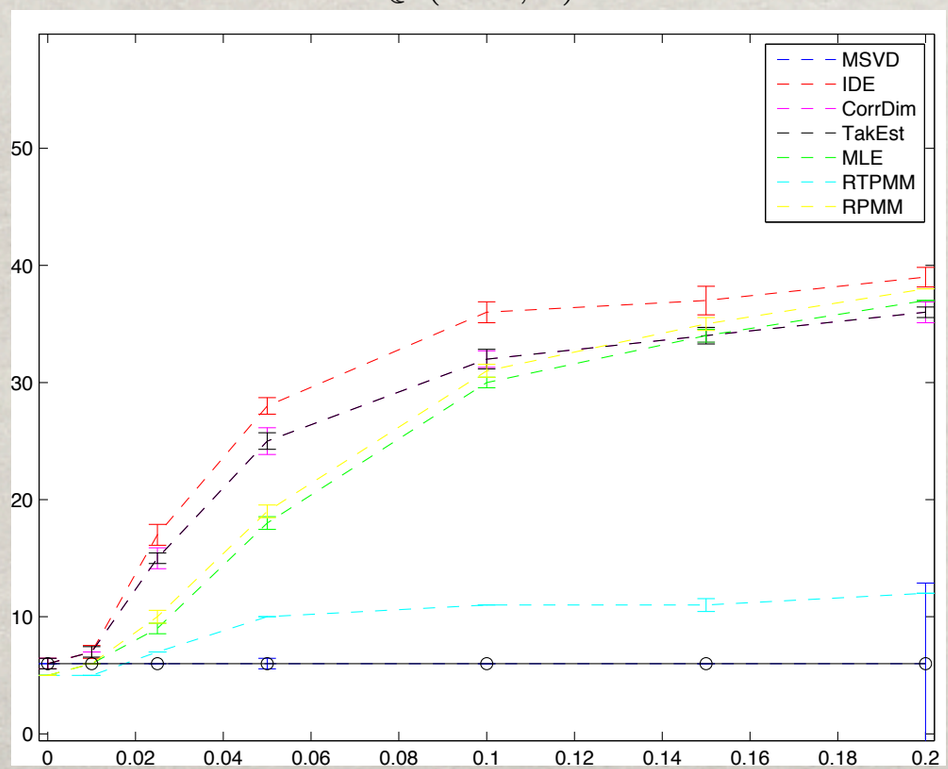
$\mathbb{Q}^6(250, \sigma)$



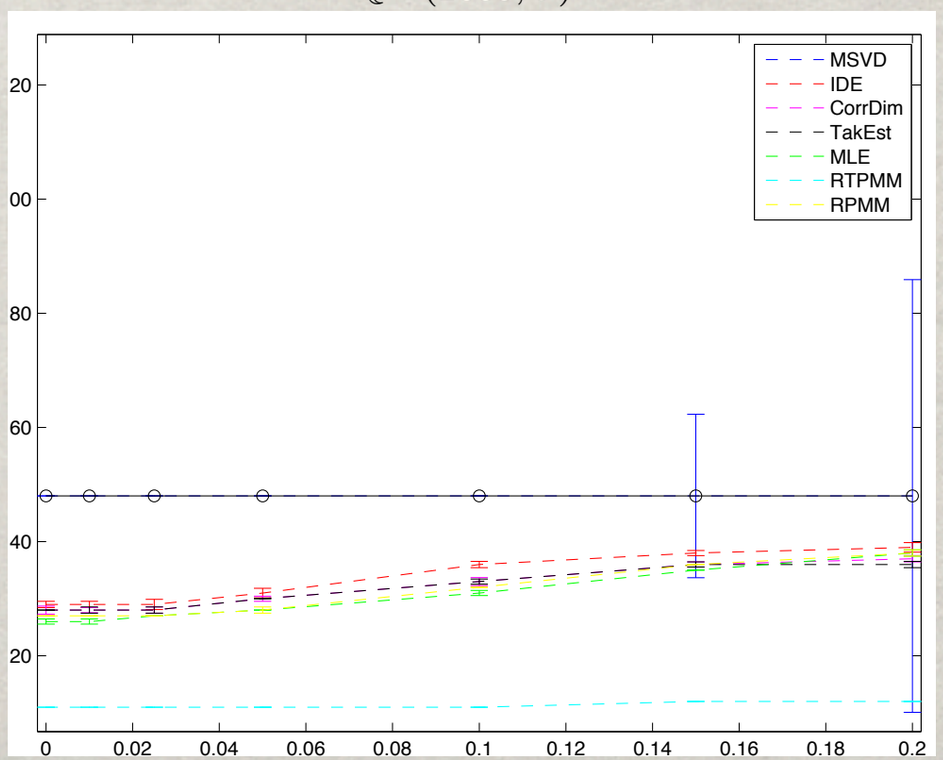
$\mathbb{Q}^{12}(1000, \sigma)$



$\mathbb{Q}^6(1000, \sigma)$

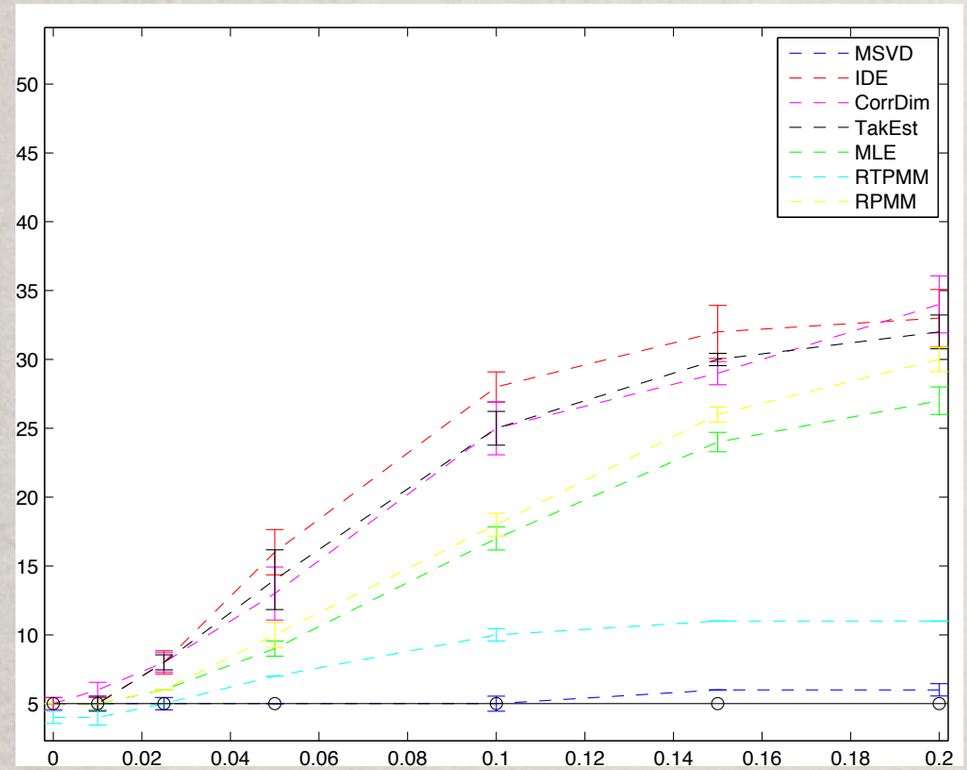


$\mathbb{Q}^{48}(2000, \sigma)$

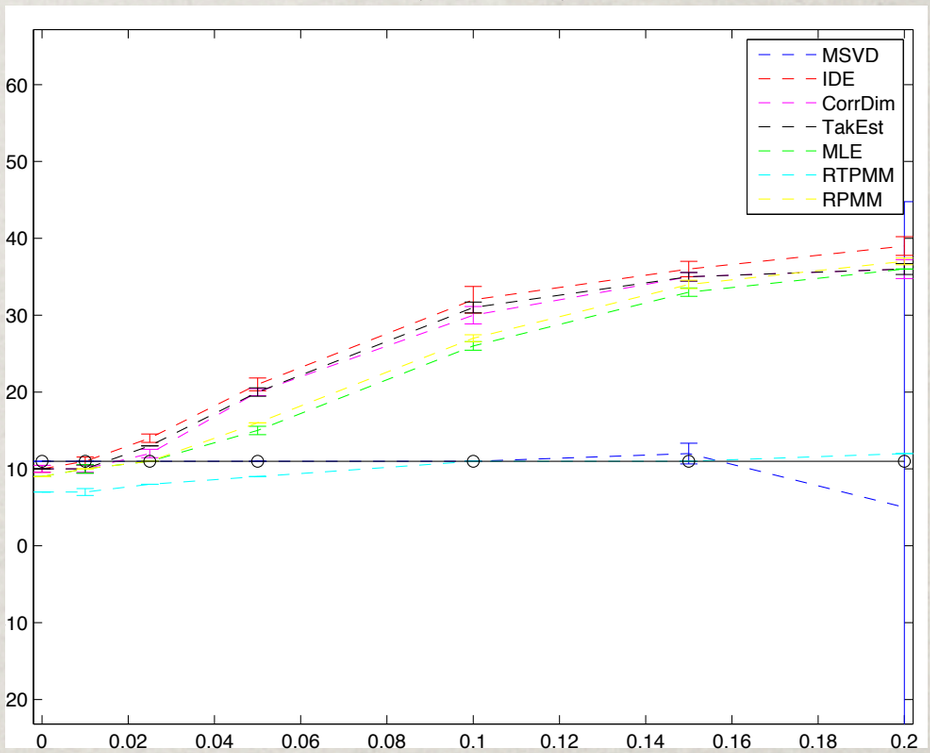


Comparison: unit sphere

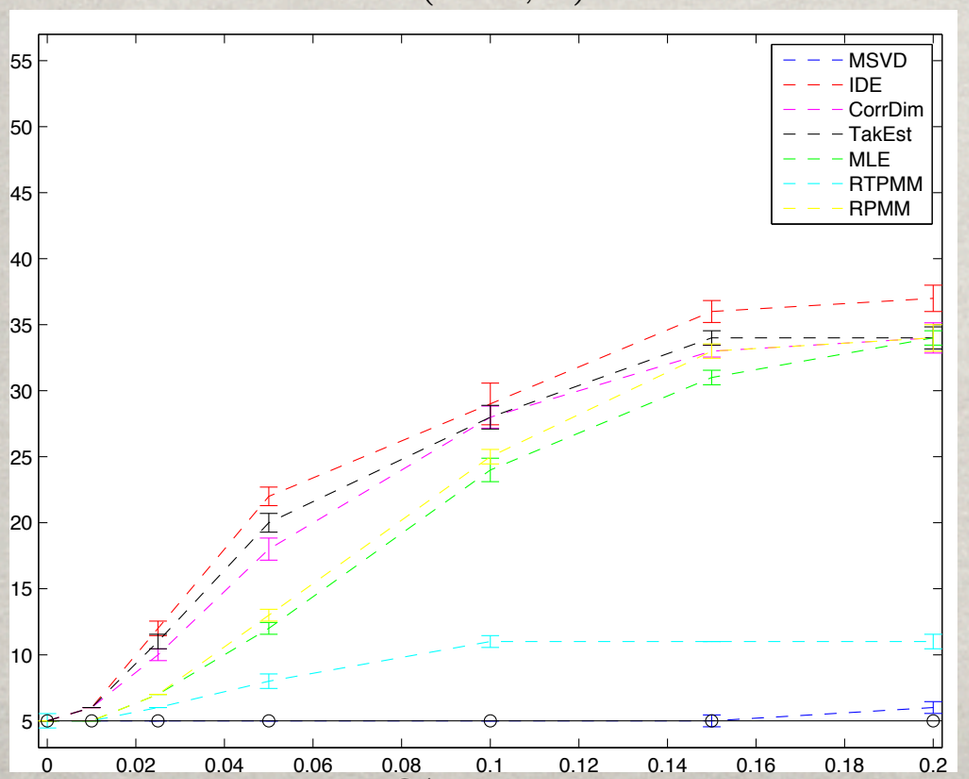
$\mathbb{S}^6(250, \sigma)$



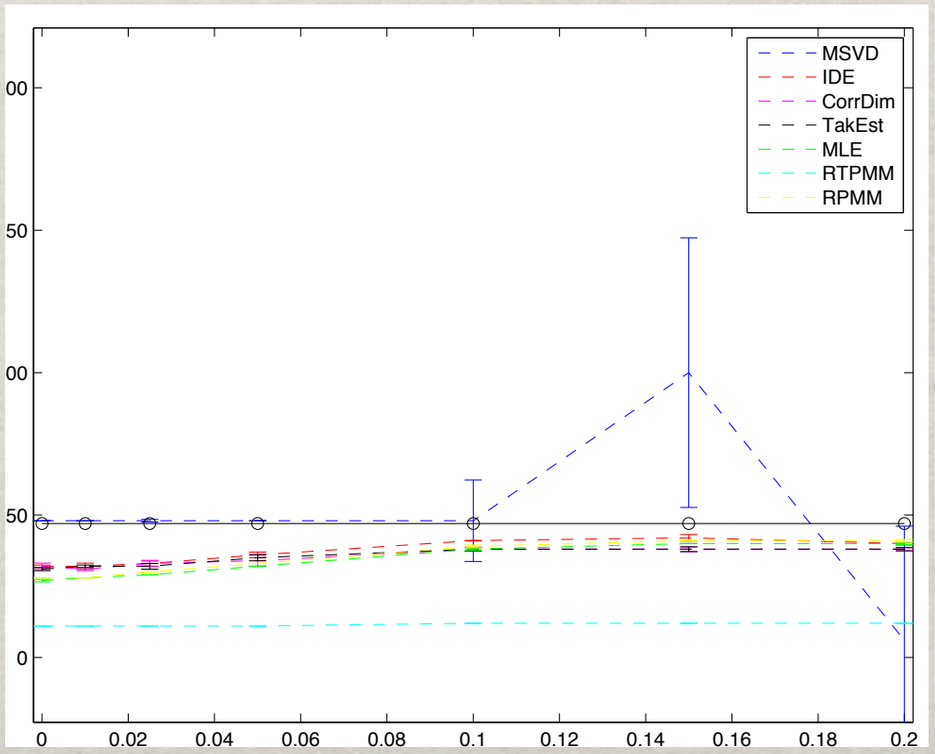
$\mathbb{S}^{12}(1000, \sigma)$



$\mathbb{S}^6(1000, \sigma)$

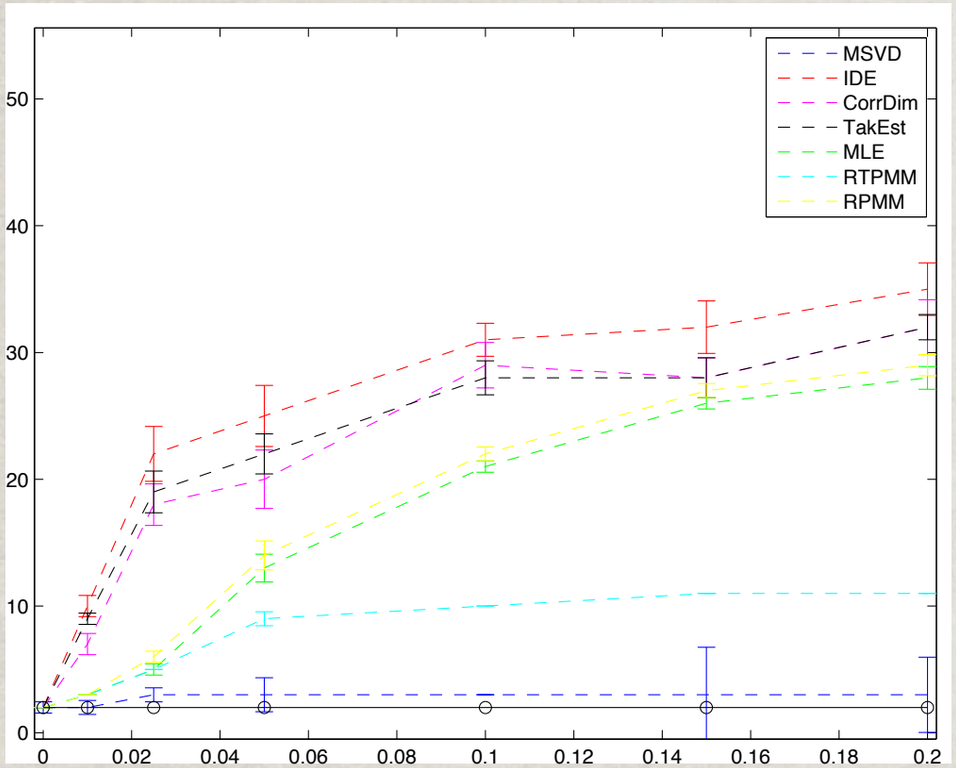


$\mathbb{S}^{48}(2000, \sigma)$

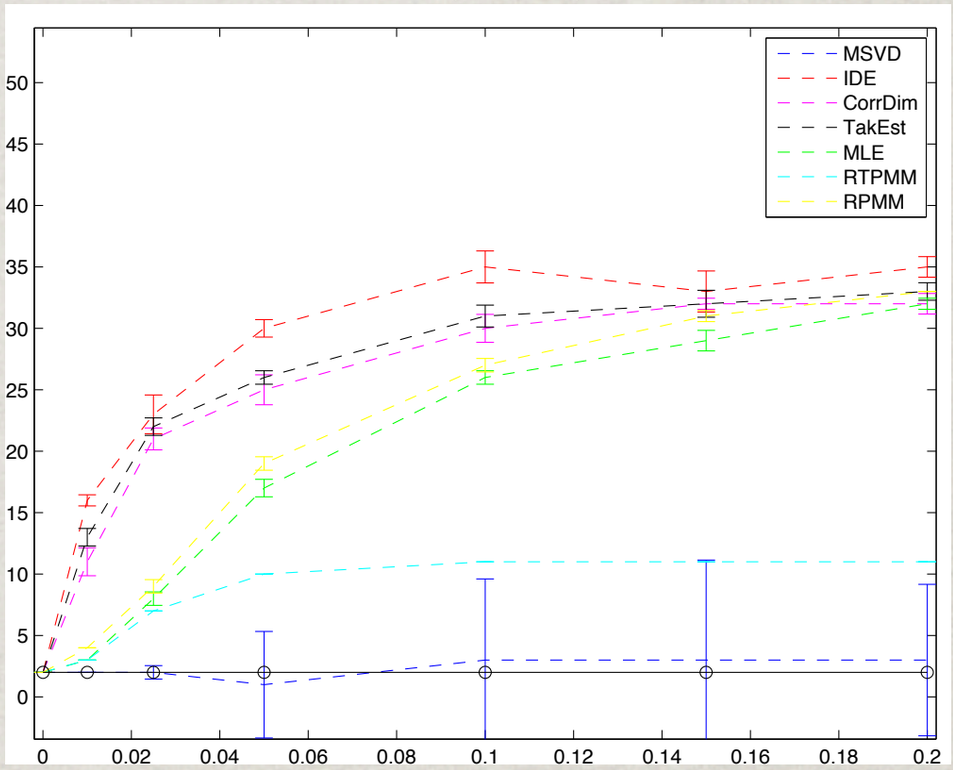


Comparison: S-shaped manifold

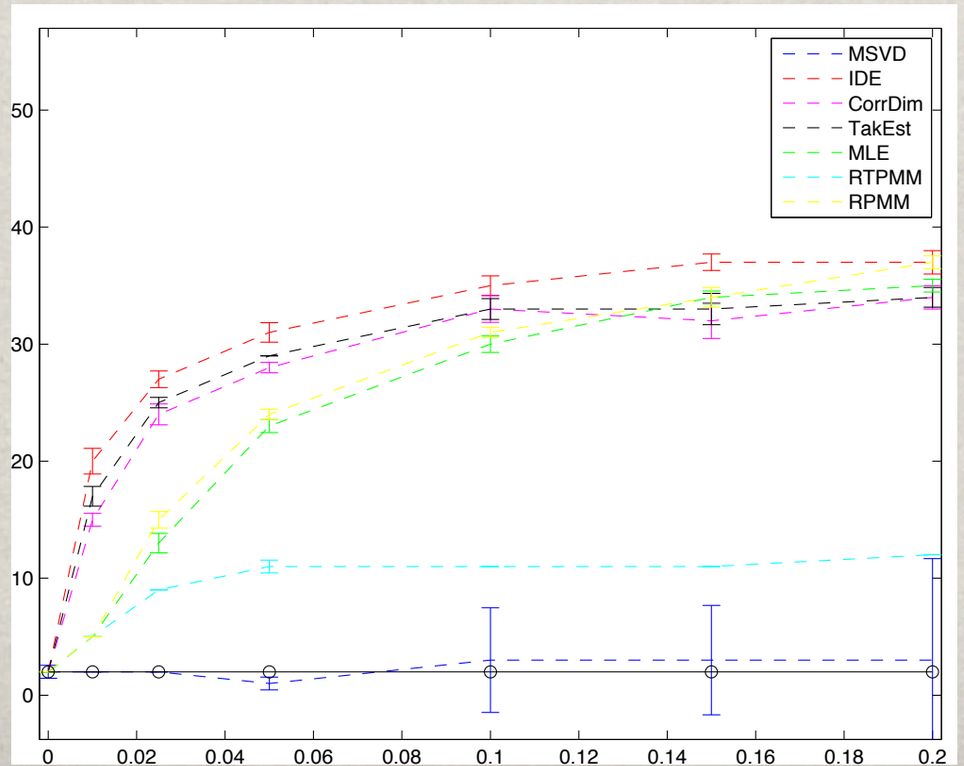
$\mathcal{Z}^{50}(200, \sigma)$



$\mathcal{Z}^{50}(500, \sigma)$



$\mathcal{Z}^{50}(1000, \sigma)$



Open problems & future dir.'s

- Faster algorithms for multiscale analysis on graphs
- Better visualization of multiscale analysis of graphs [E. Monson, R. Brady]
- Implementation and use in visualization of algorithms derived from eigenfunction/heat kernel embeddings
- Generalization to data sets with varying dimensionality [A. Little, J. Lee]
- Applications to real world data sets [A. Little, M. Crosskey; C. Clementi; L. Rosasco]
- Towards a toolbox of highly robust geometric analysis tools for data sets [A. Little, G. Chen].
- Dynamic graphs [K. Balachandrian, J. Lee]

Collaborators: E. Monson, R. Brady (Duke C.S.); R. Coifman (Math, Yale), P.W. Jones (Math, Yale); R. Schul (Math, Stonybrook); A. V. Little, K. Balachandrian (Math grad, Duke), J. Lee (Math undergrad, Duke); L. Rosasco (CS, MIT and Universita' di Genova); C. Clementi (Chem., Rice); S. Mukherjee (Stat, Duke); J. Guinney (Comp. Bio. grad, Duke); P. Febbo (Med., Duke).
Funding: NSF, ONR, Sloan Foundation, Duke.

www.math.duke.edu/~mauro