

# Geometric analysis of molecular dynamics data, diffusion geometry and reaction coordinates

Mauro Maggioni

Department of Mathematics and Computer Science  
Duke University

I.P.A.M. - 4/12/2011

Joint work: G. Chen, A. Little (Duke)

C. Clementi, M.A. Rohrdanz and W. Zheng (Rice)

Partial support: DARPA, NSF, ONR, Sloan

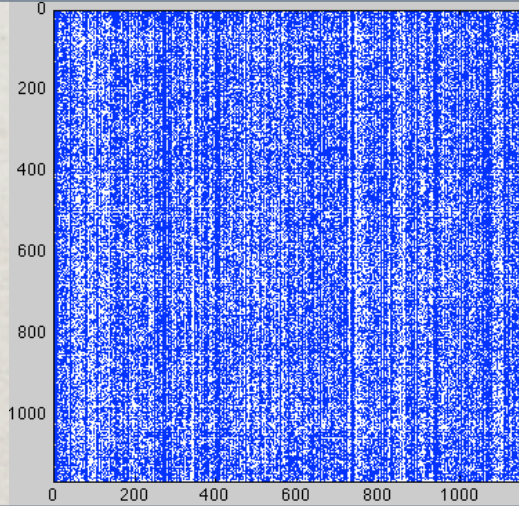
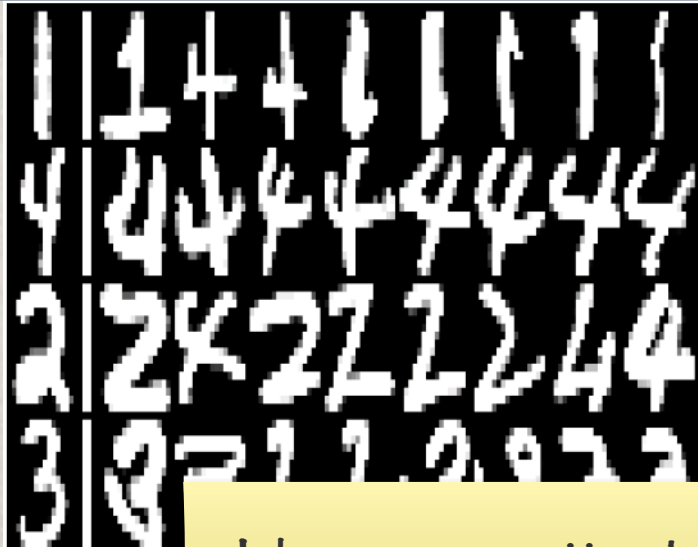




# Data Sets in High-Dimensions

*A deluge of data:* documents, customer databases, images, social network transactions, gene arrays, sensor networks, financial transactions...

*Data set:* often  $X \subset \mathbb{R}^D$ ,  $D$  very large ( $10^2 - 10^8$ ).

Data	Picture	Problems
Approx. 1000 articles from ScienceNews. Representation: a document-term matrix, with about 1000 terms.		Automatically sort articles into categories, given only a small set labeled by experts.  Navigate the library.
A database of ~60000 grayscale 28x28 images of handwritten digits 0-9.		Automatically recognize digits (e.g. for ZIP codes, checks, etc...)

I show non-zero pattern in the word-doc matrix, I promise frequencies wouldn't help!



# A fragmented landscape

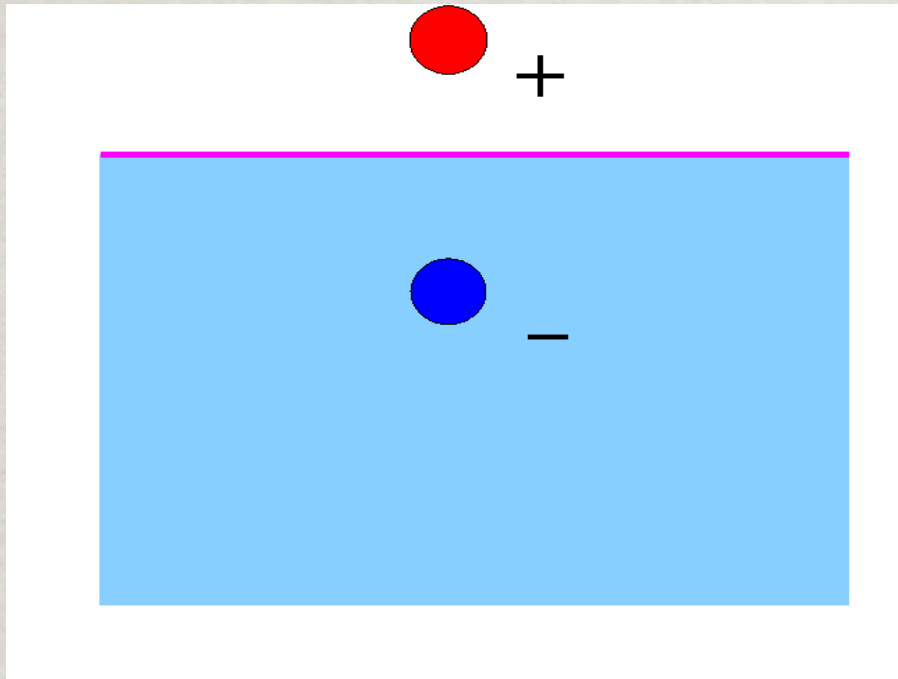
- . Problems arise in many applications, and research fields (computer science, engineering, applied mathematics, statistics, biology, ...).
- . No single approach will be optimal in all cases, but there are common fundamental questions and common promising approaches.
- . Data model: set of samples from a distribution in  $\mathbb{R}^D$ ; set of vertices of a graph.
- . **Geometric problems**: which geometric properties does the data have? Low intrinsic dimension, manifold-like, easily partitioned into clusters, etc...
- . **Function approximation problems**: we are interested in learning and predicting certain observables from data. We need to approximate a function, defined on the data, and thereby with possibly high-dimensional domain.
- . One may try to **fuse the two problems** above.

comment on where machine learning sits



# Learning & Geometry

**The Geometric Basis of Semi-supervised Learning.** V. Sindhwani, M. Belkin, and P. Niyogi, in *Semi-supervised Learning*, (Chapelle, Schoelkopf, Zien: editors), MIT Press, 2006.



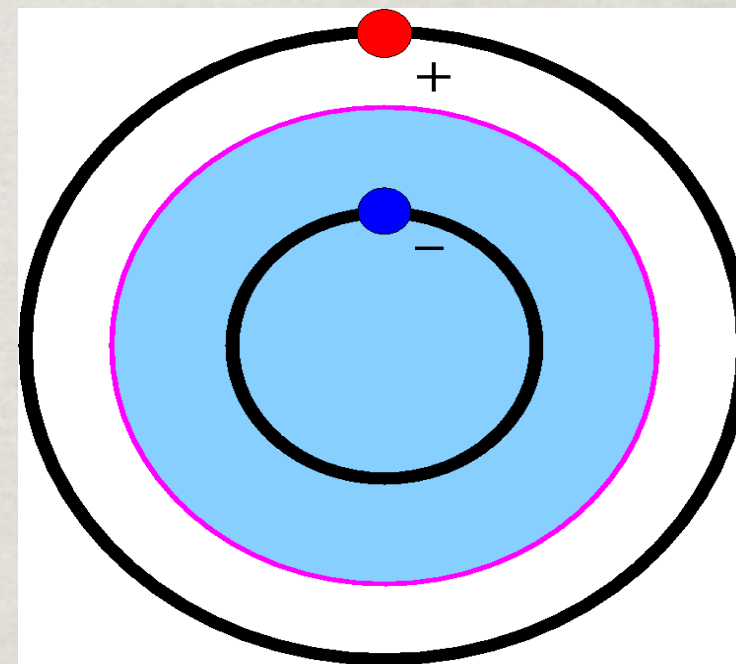
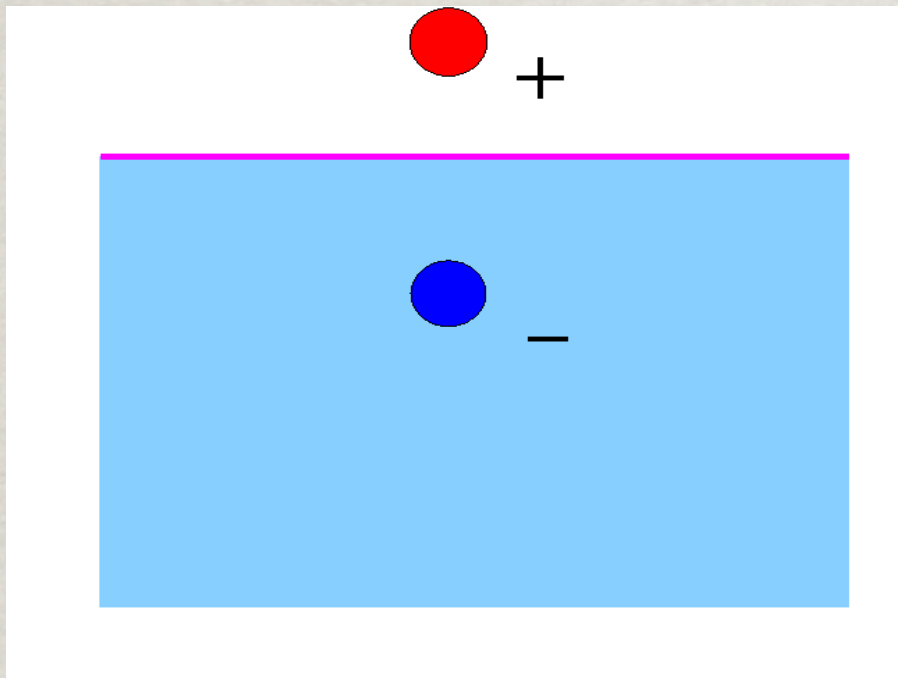
Tie geometry and learning of functions.  
This was one great insight of University of Chicago very own Partha Niyogi, whom I am sure several of us sorely miss.  
I borrowed these figures from one of his papers, where with his usual clarity he emphasizes how important ties between geometry and learning may be.

Lots of work in the past 10 years in the machine learning, statistical dimensionality reduction, topological data analysis.



# Learning & Geometry

**The Geometric Basis of Semi-supervised Learning.** V. Sindhwani, M. Belkin, and P. Niyogi, in *Semi-supervised Learning*, (Chapelle, Schoelkopf, Zien: editors), MIT Press, 2006.



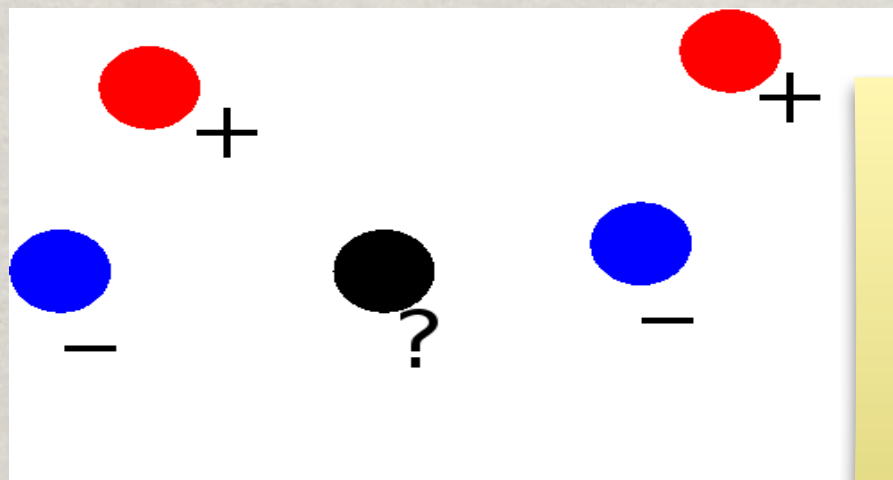
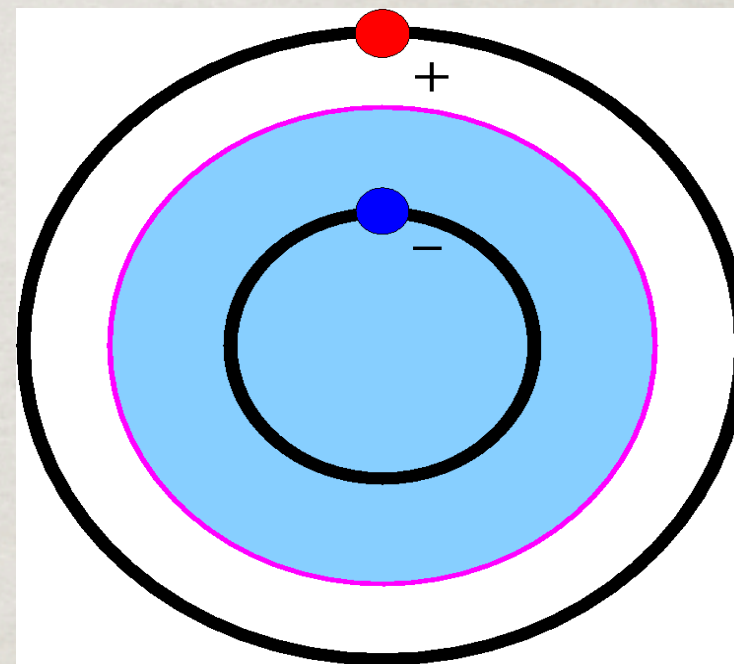
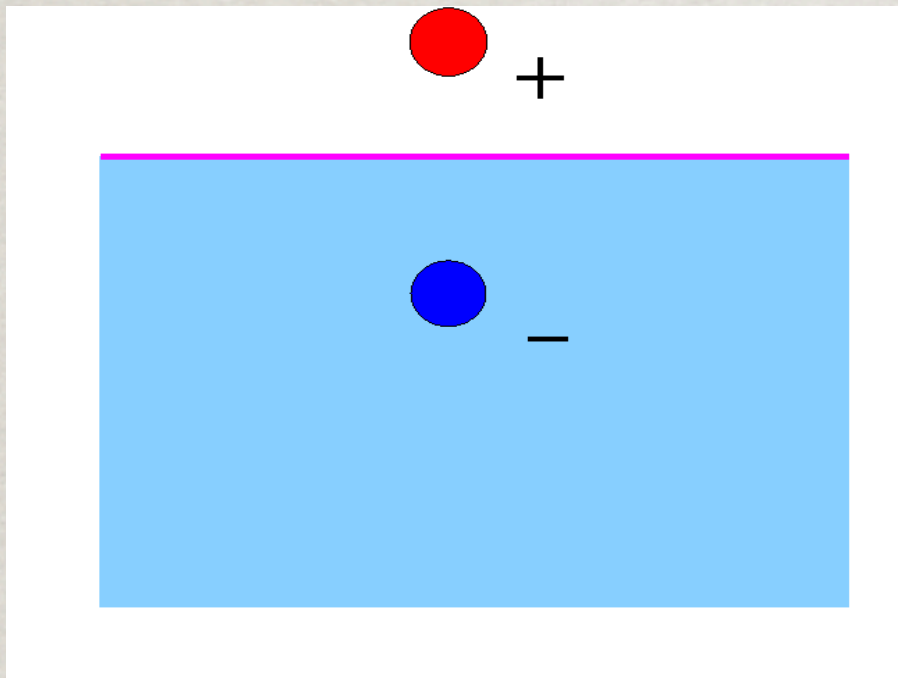
Tie geometry and learning of functions.  
This was one great insight of University of Chicago very own Partha Niyogi, whom I am sure several of us sorely miss.  
I borrowed these figures from one of his papers, where with his usual clarity he emphasizes how important ties between geometry and learning may be.

Lots of work in the past 10 years in the machine learning, statistical dimensionality reduction, topological data analysis.



# Learning & Geometry

**The Geometric Basis of Semi-supervised Learning.** V. Sindhwani, M. Belkin, and P. Niyogi, in *Semi-supervised Learning*, (Chapelle, Schoelkopf, Zien: editors), MIT Press, 2006.



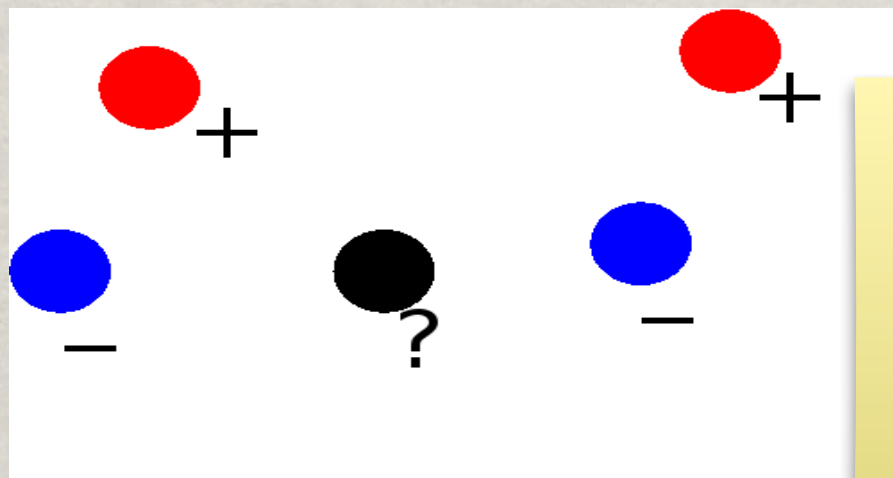
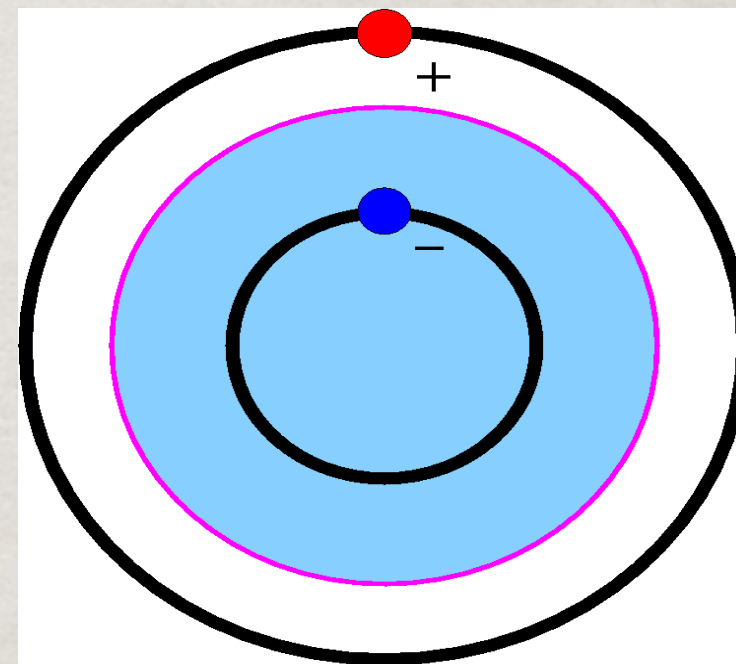
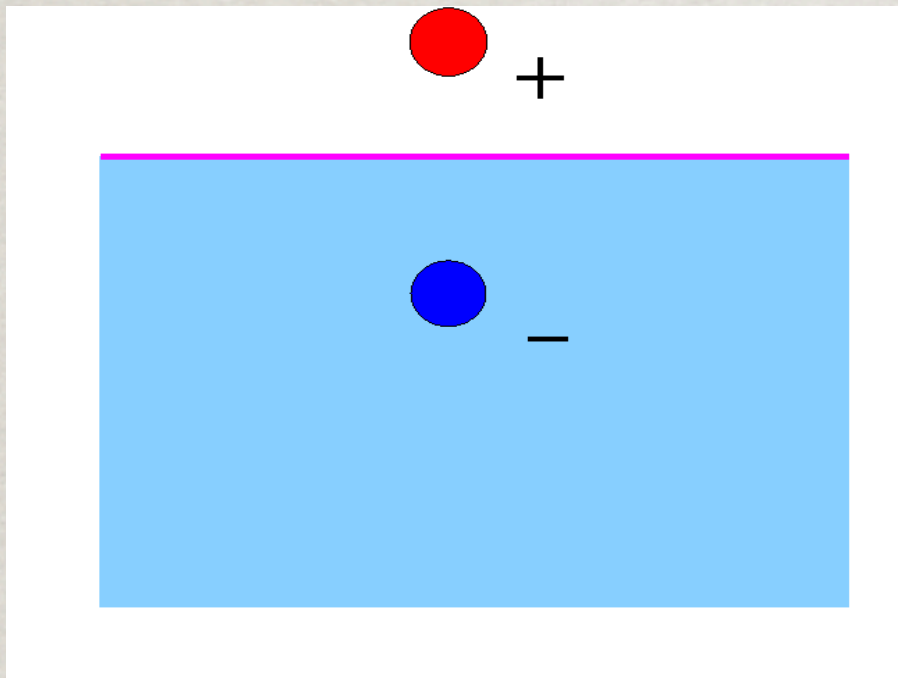
Tie geometry and learning of functions.  
This was one great insight of University of Chicago very own Partha Niyogi, whom I am sure several of us sorely miss.  
I borrowed these figures from one of his papers, where with his usual clarity he emphasizes how important ties between geometry and learning may be.

Lots of work in the past 10 years in the machine learning, statistical dimensionality reduction, topological data analysis.



# Learning & Geometry

**The Geometric Basis of Semi-supervised Learning.** V. Sindhwani, M. Belkin, and P. Niyogi, in *Semi-supervised Learning*, (Chapelle, Schoelkopf, Zien: editors), MIT Press, 2006.



Tie geometry and learning of functions.  
This was one great insight of University of Chicago very own Partha Niyogi, whom I am sure several of us sorely miss.  
I borrowed these figures from one of his papers, where with his usual clarity he emphasizes how important ties between geometry and learning may be.

Lots of work in the past 10 years in the machine learning, statistical dimensionality reduction, topological data analysis.



# Random walks on data and graphs

Given:

Joint with R. Coifman and S. Lafon

- . **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

$$\underbrace{P = D^{-1}W}_{\text{random walk}}, \quad \underbrace{T = D^{-\frac{1}{2}}WD^{-\frac{1}{2}}}_{\text{symm. "random walk"}}, \quad \underbrace{L = I - T}_{\text{norm. Laplacian}}, \quad \underbrace{H = e^{-tL}}_{\text{Heat kernel}}$$

One may start doing analysis and measure smoothness:

$$\langle Lf, f \rangle = \sum_x \sum_{y \sim x} W(x, y) \left( \frac{f(x)}{\sqrt{d_x}} - \frac{f(y)}{\sqrt{d_y}} \right)^2 \sim \int |\nabla f|^2 dW$$



# 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  $n \rightarrow \infty$  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  $\mathcal{M}$ .



# BASIC NORMS ON GRAPHS

Any function  $f : G \rightarrow \mathbb{R}$  is a vector in  $\mathbb{R}^N$ . Euclidean norm and inner product:

$$\|f\|_2^2 = \sum_{x \in G} |f(x)|^2 d(x) \quad , \quad \langle f, g \rangle = \sum_{x \in G} f(x)g(x)d(x)$$

where  $d(x) = \sum_{y \sim x} W(x, y)$ .

Other choices are possible

A Laplacian  $L$  allows to introduce a notion of smoothness

$$\langle Lf, f \rangle = \sum_x \sum_{y \sim x} W(x, y) \left( \frac{f(x)}{\sqrt{d_x}} - \frac{f(y)}{\sqrt{d_y}} \right)^2 \sim \int_{\text{edges}} |\nabla f|^2 dW$$

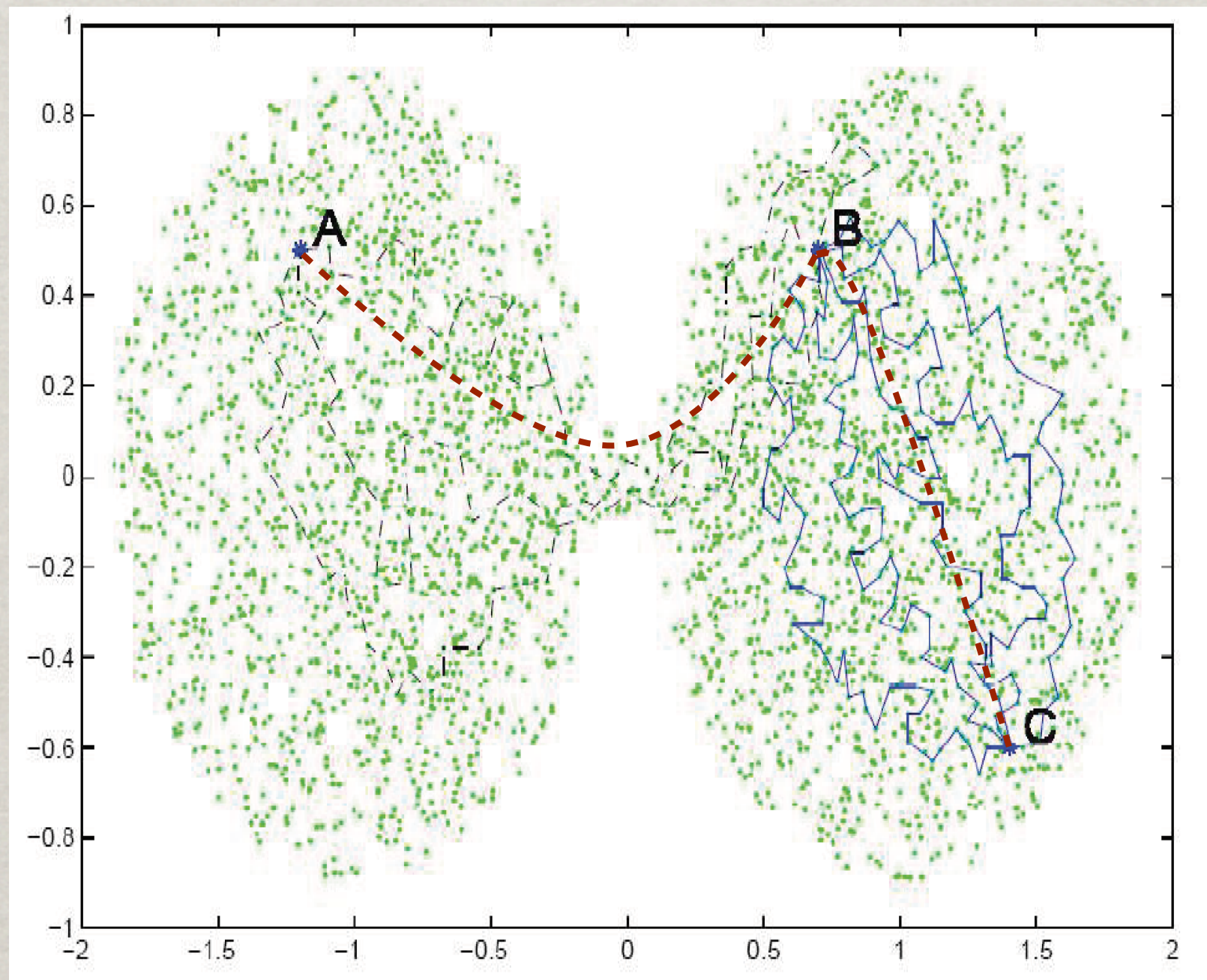
Moreover,

$$\lambda_i(L) = \min_{f \perp \langle \varphi_1, \dots, \varphi_{i-1} \rangle} \frac{\langle Lf, f \rangle}{\|f\|^2}$$



# Diffusion Distances, I

In some cases the geodesic distances  $d_{\mathcal{M}}$  may not capture interesting geometric information. For example here  $d_{\mathcal{M}}(A, B) \sim d_{\mathcal{M}}(B, C)$ . Can we define a new distance that may capture this type of geometric characteristic?




Picture courtesy of S. Lafon



# Diffusion Distances, II

We may use random walks, and, for  $t > 0$ , define



$$d^{(t)}(x, y) = \|T^t(x, \cdot) - T^t(y, \cdot)\|_{L^2(G)} = \sqrt{\sum_{z \in G} |T^t(x, z) - T^t(y, z)|^2}$$

$$\stackrel{T\varphi_i = \lambda_i\varphi_i}{=} \sqrt{\sum_{i=1}^{+\infty} \lambda_i^{2t} (\varphi_i(x) - \varphi_i(y))^2}$$

$$\sim \left\| \left( \lambda_i^t \varphi_i(x) \right)_{i=1}^m - \left( \lambda_i^t \varphi_i(y) \right)_{i=1}^m \right\|_{\mathbb{R}^m}$$

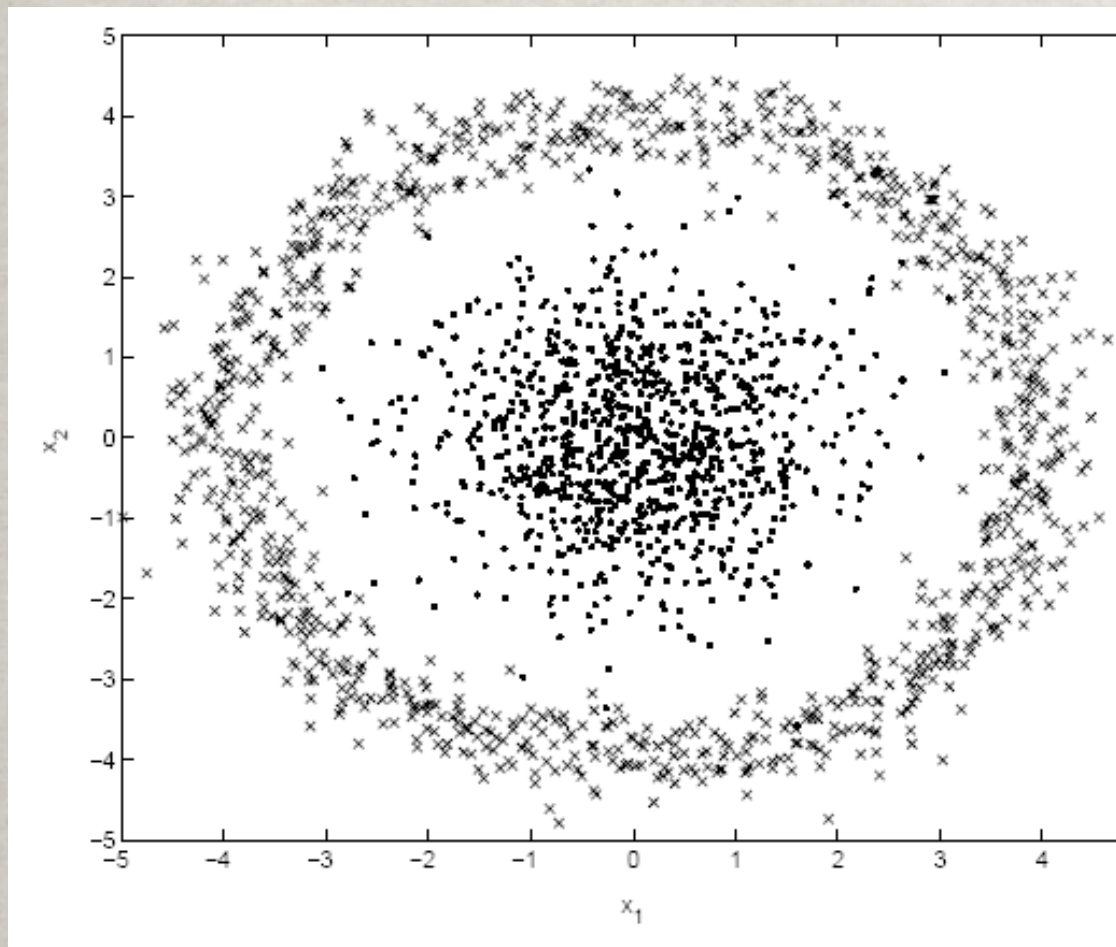
Therefore  $\Phi_m^{(t)}$  defined by  $\Phi_m^{(t)}(x) = \left( \lambda_i^t \varphi_i(x) \right)_{i=1}^m$  satisfies

$$\left\| \Phi_m^{(t)}(x) - \Phi_m^{(t)}(y) \right\|_{\mathbb{R}^m} \sim d^{(t)}(x, y),$$

at least for  $t$  large and  $m$  large.



# Spectral Clustering in one slide



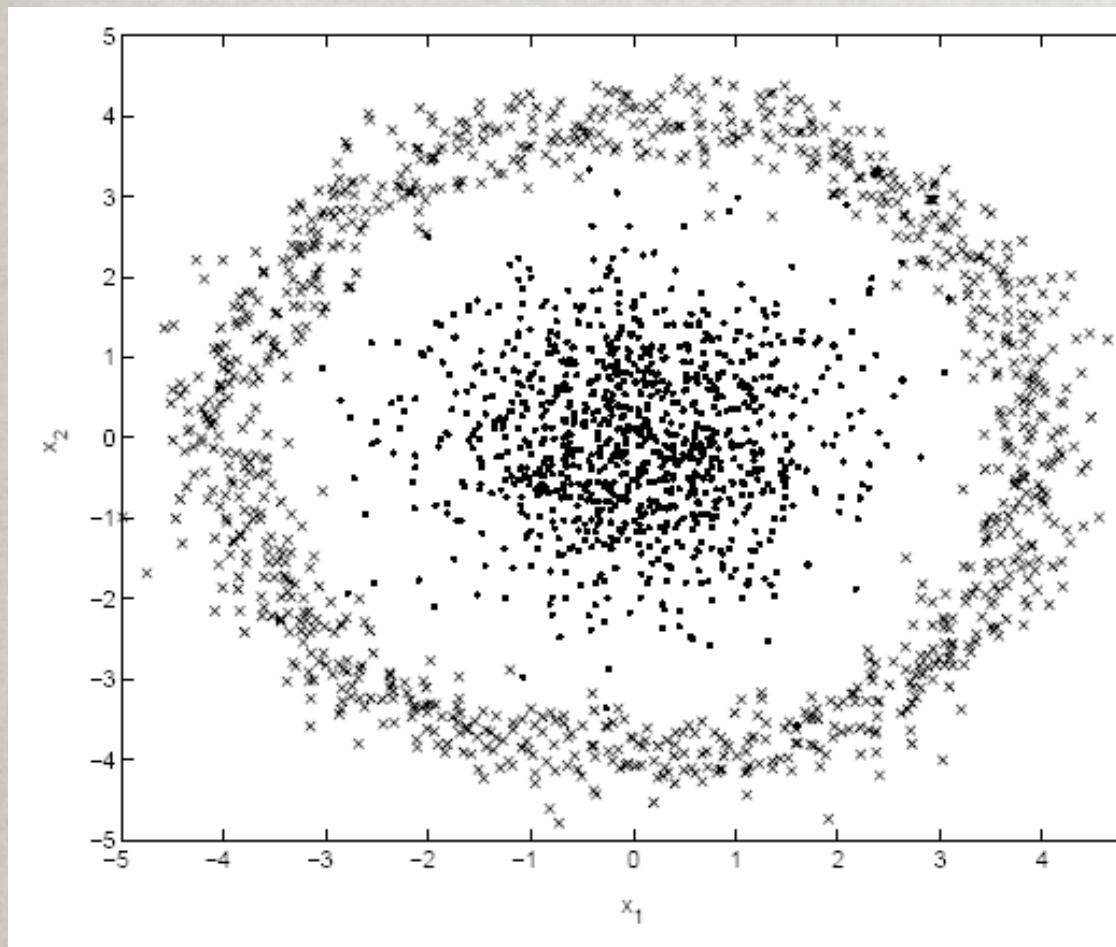
## Original space

Every point is connected to its 5 nearest neighbors, to obtain a graph.



# Spectral Clustering in one slide

$$x \mapsto (\phi_2(x), \phi_3(x))$$



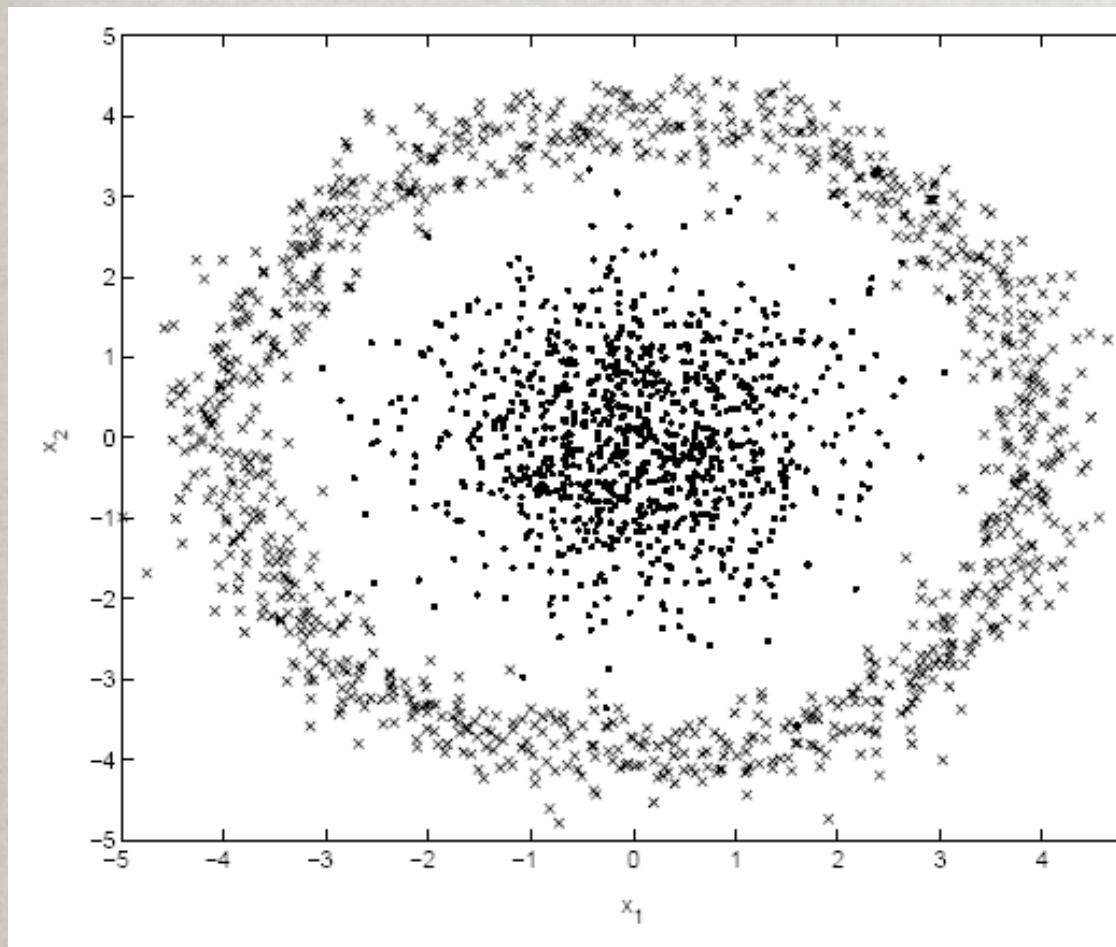
## Original space

Every point is connected to its 5 nearest neighbors, to obtain a graph.



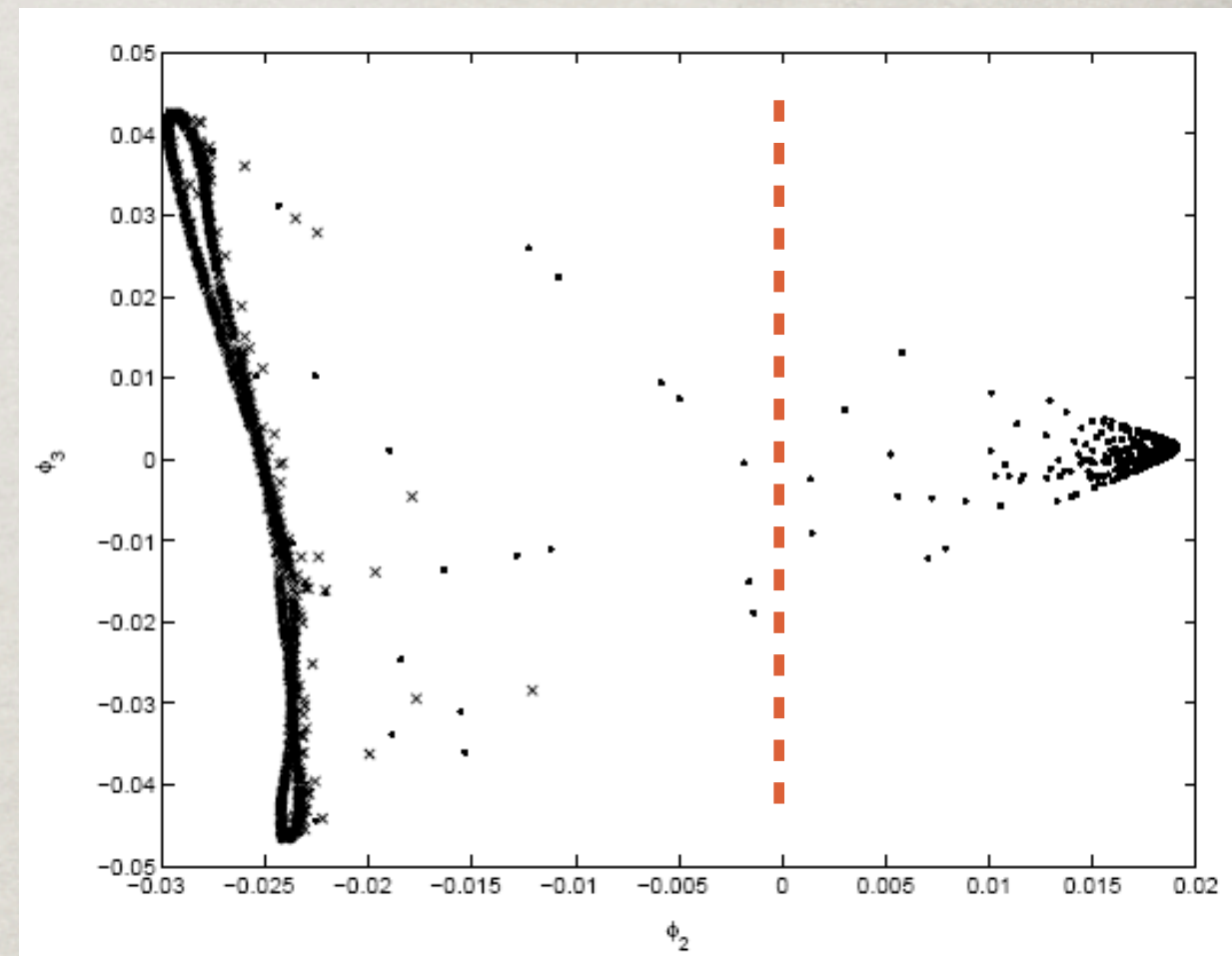
# Spectral Clustering in one slide

$$x \mapsto (\phi_2(x), \phi_3(x))$$



## Original space

Every point is connected to its 5 nearest neighbors, to obtain a graph.



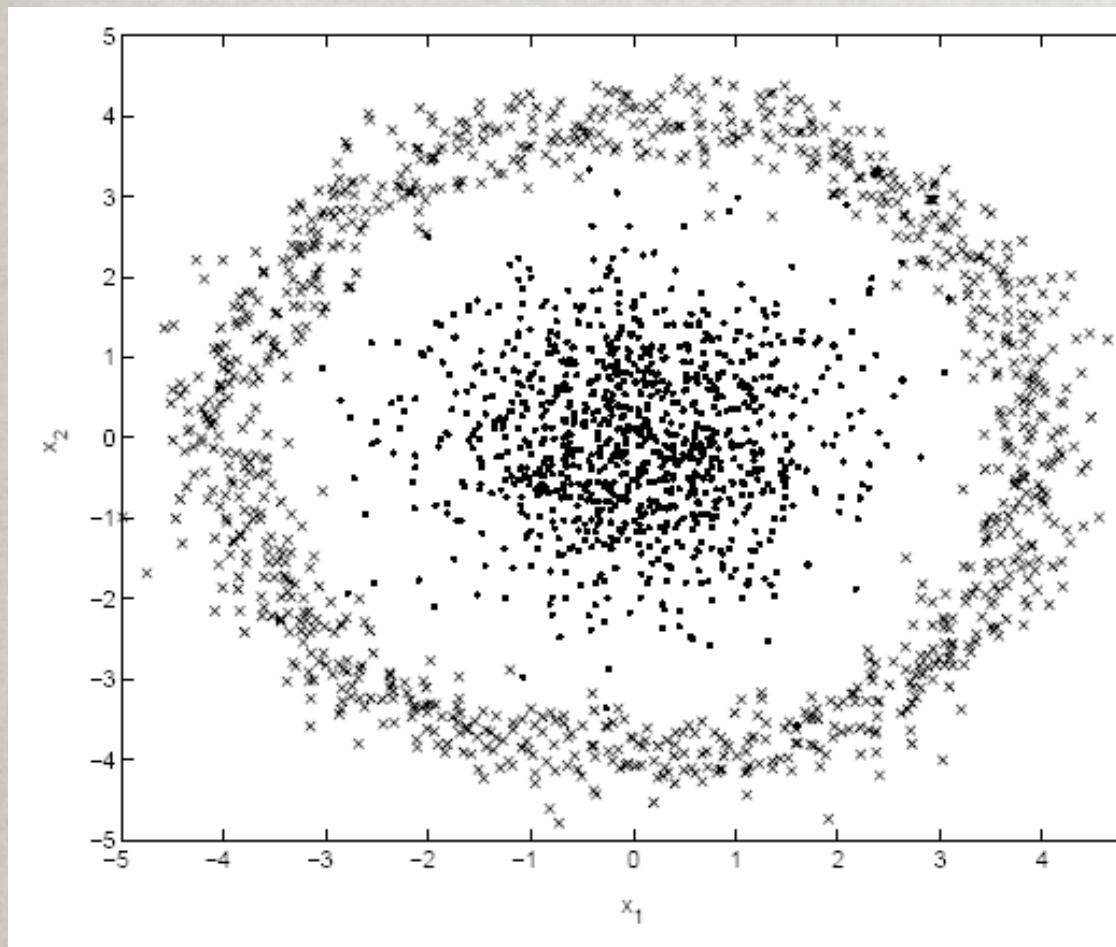
## Diffusion space

$\phi_2 = 0$  is a good cut!



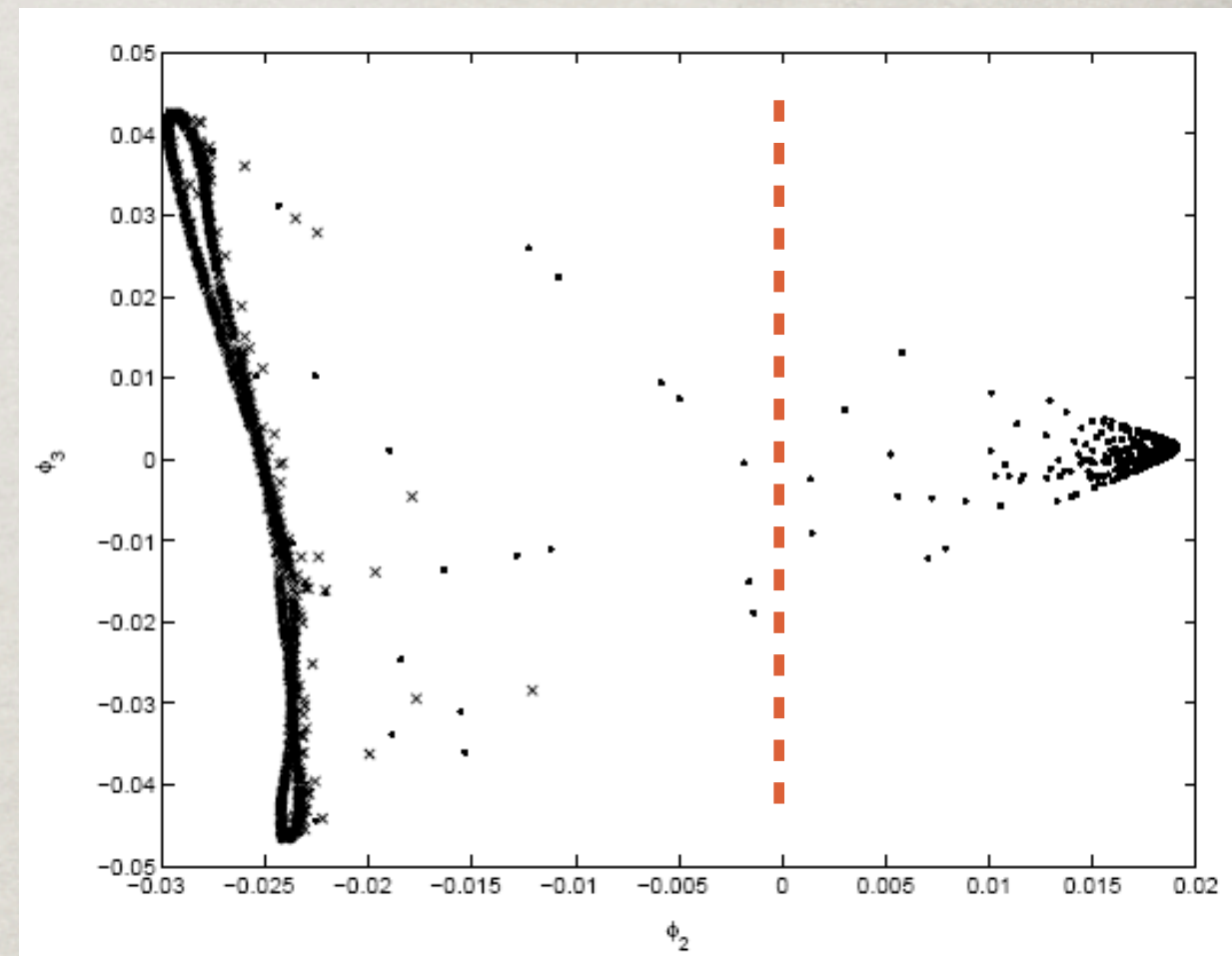
# Spectral Clustering in one slide

$$x \mapsto (\phi_2(x), \phi_3(x))$$



## Original space

Every point is connected to its 5 nearest neighbors, to obtain a graph.



## Diffusion space

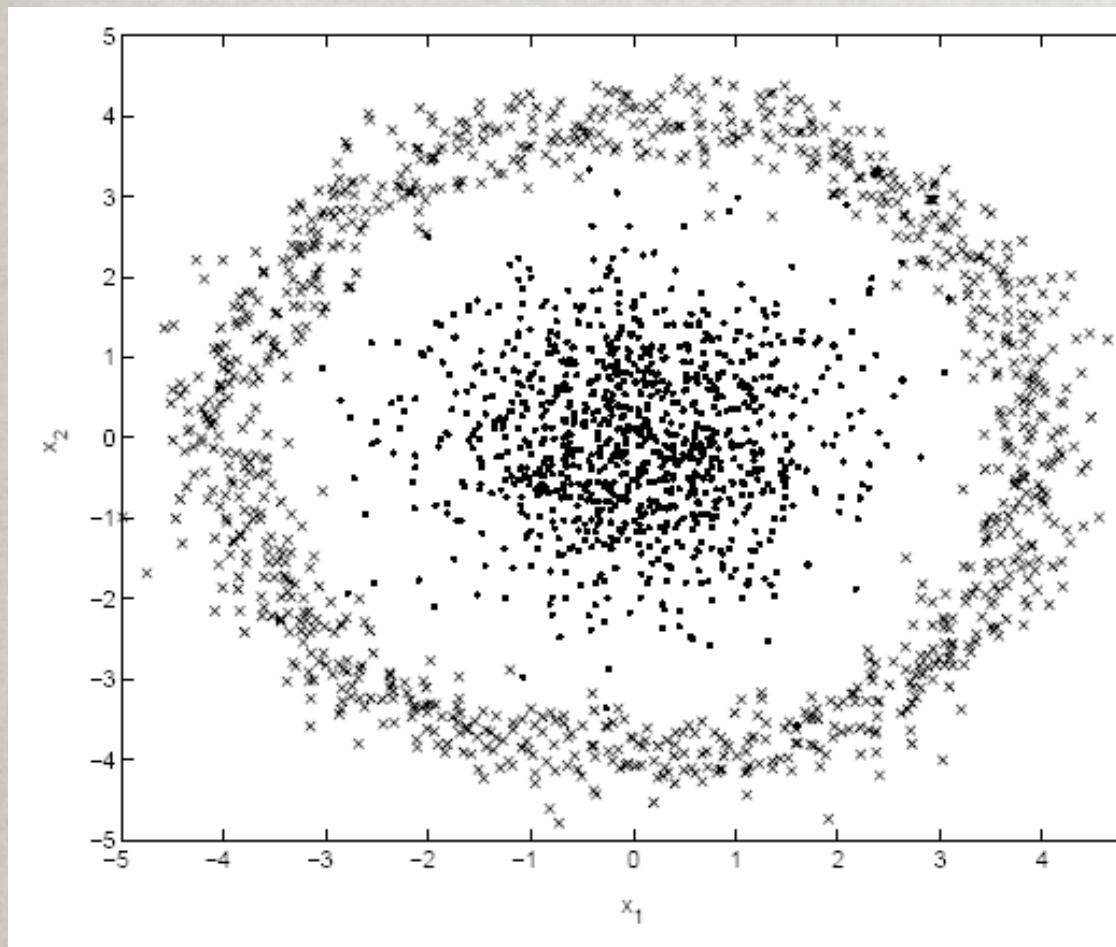
$\phi_2 = 0$  is a good cut!

$$\langle Lf, f \rangle = \sum_x \sum_{y \sim x} W(x, y) \left( \frac{f(x)}{\sqrt{d_x}} - \frac{f(y)}{\sqrt{d_y}} \right)^2$$



# Spectral Clustering in one slide

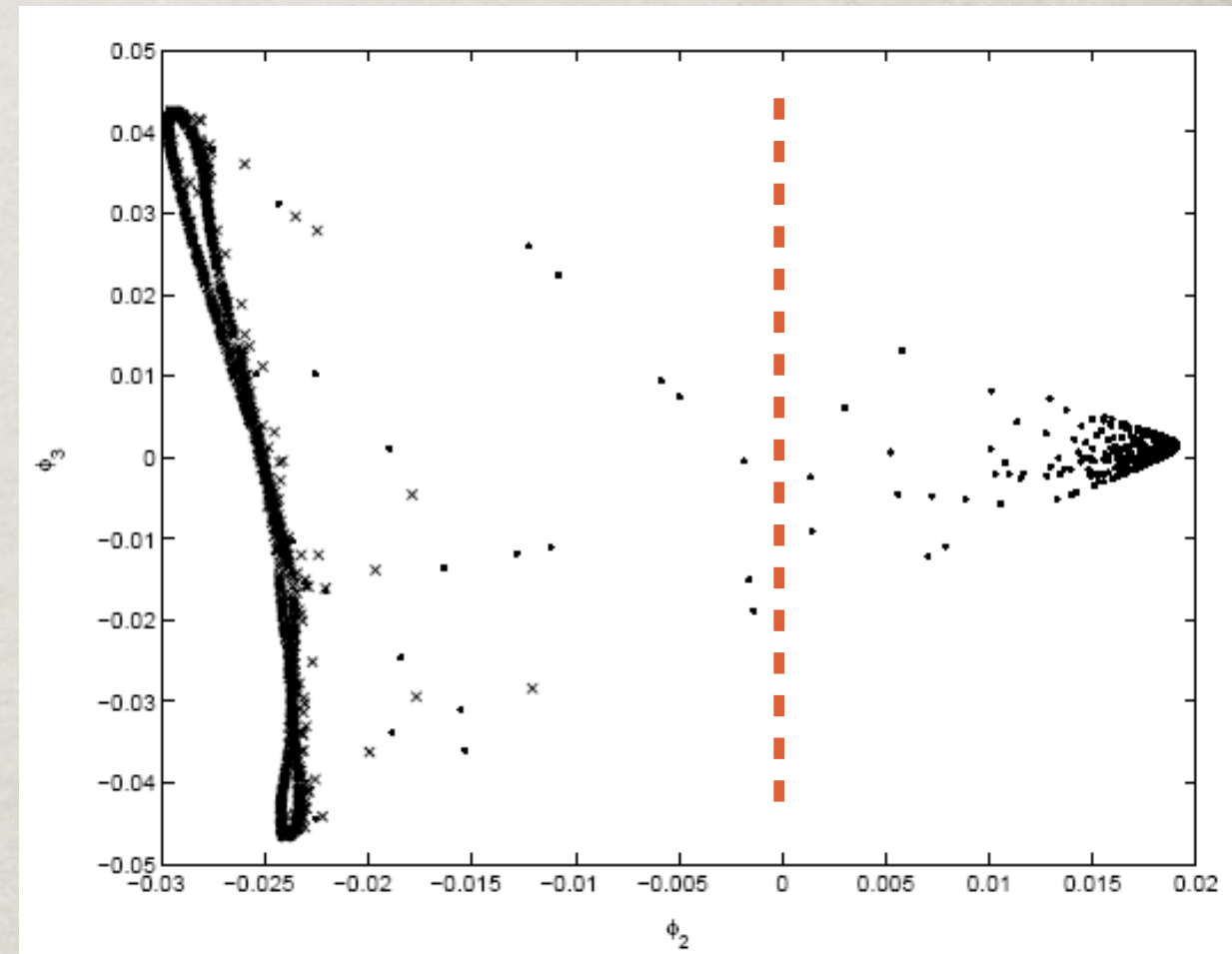
$$x \mapsto (\phi_2(x), \phi_3(x))$$



## Original space

Every point is connected to its 5 nearest neighbors, to obtain a graph.

Flexibility + robustness



## Diffusion space

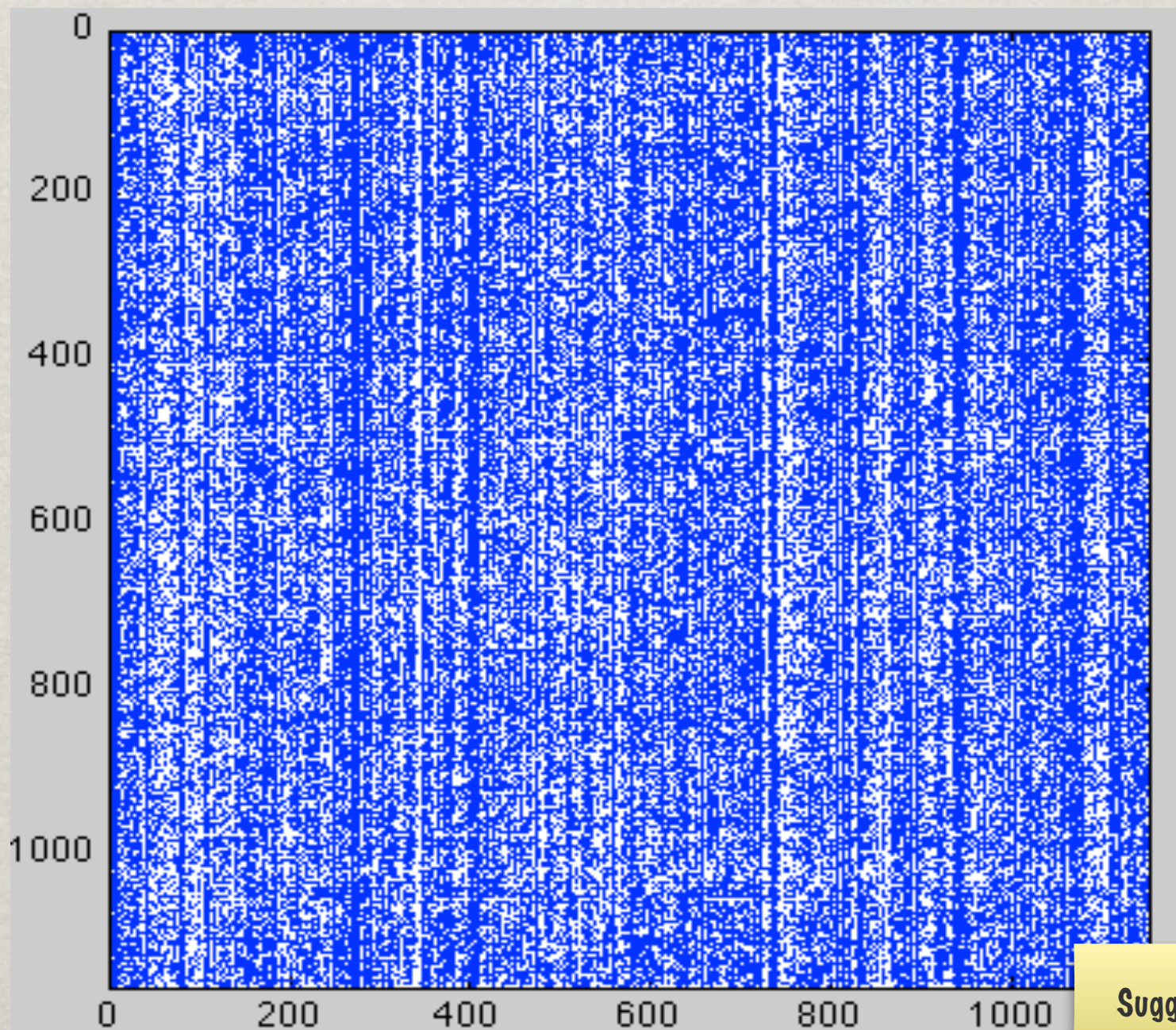
$\phi_2 = 0$  is a good cut!

$$\langle Lf, f \rangle = \sum_x \sum_{y \sim x} W(x, y) \left( \frac{f(x)}{\sqrt{d_x}} - \frac{f(y)}{\sqrt{d_y}} \right)^2$$



# Example: 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. Point cloud of 1100 points in  $\mathbb{R}^{1000}$ .

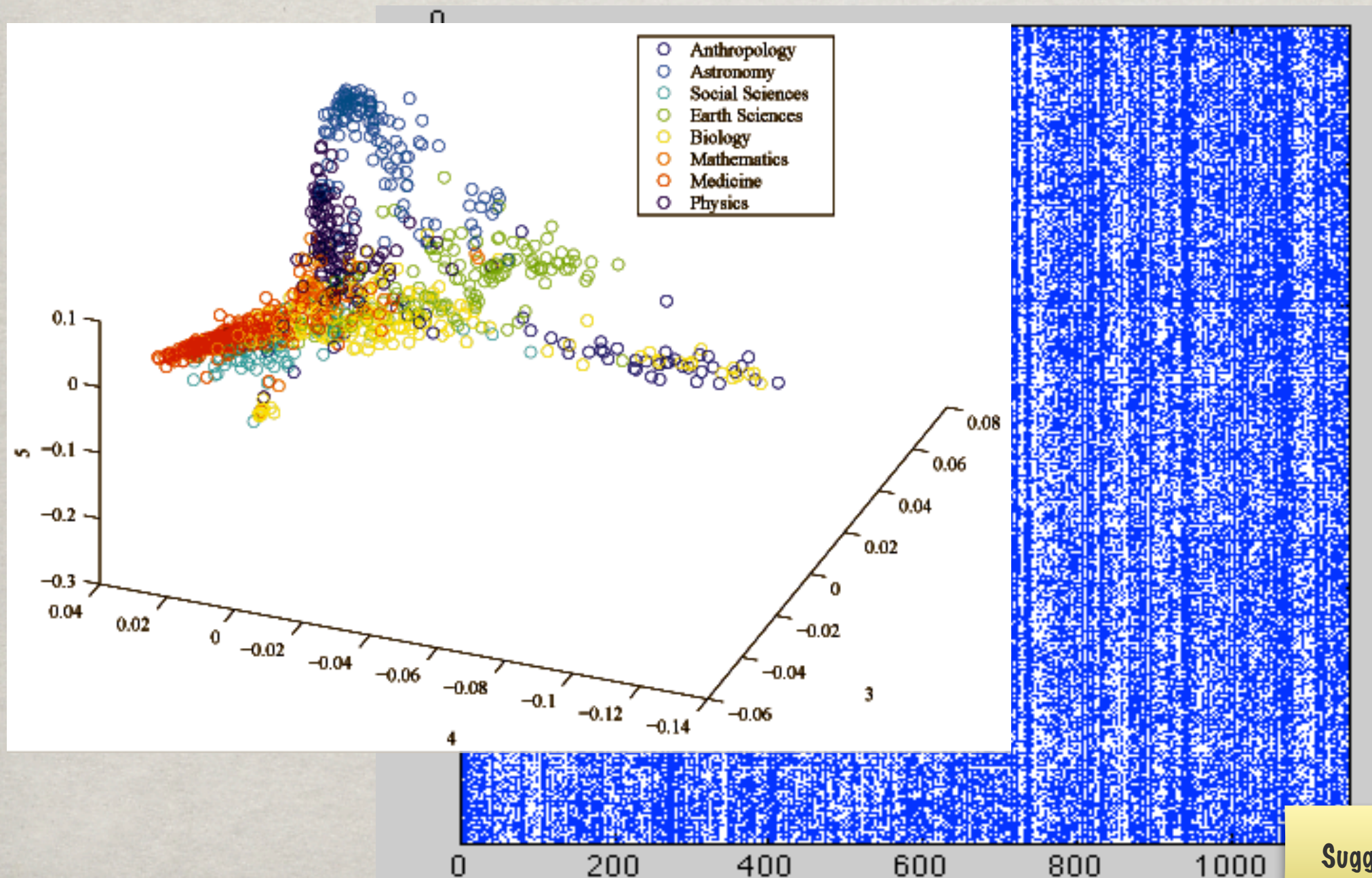


Suggest low-intrinsic dimension



# Example: 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. Point cloud of 1100 points in  $\mathbb{R}^{1000}$ .

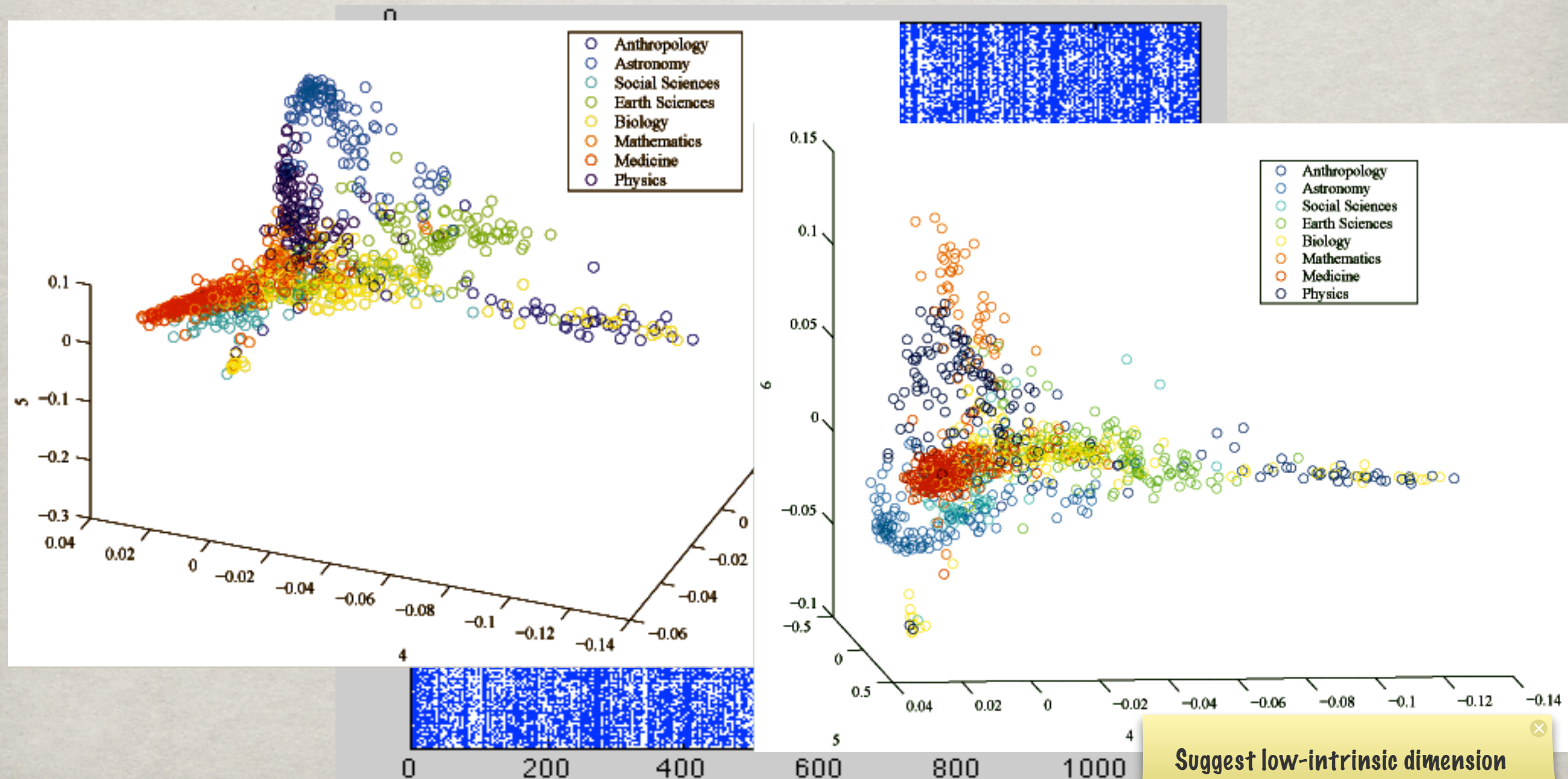


Suggest low-intrinsic dimension



# Example: 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. Point cloud of 1100 points in  $\mathbb{R}^{1000}$ .





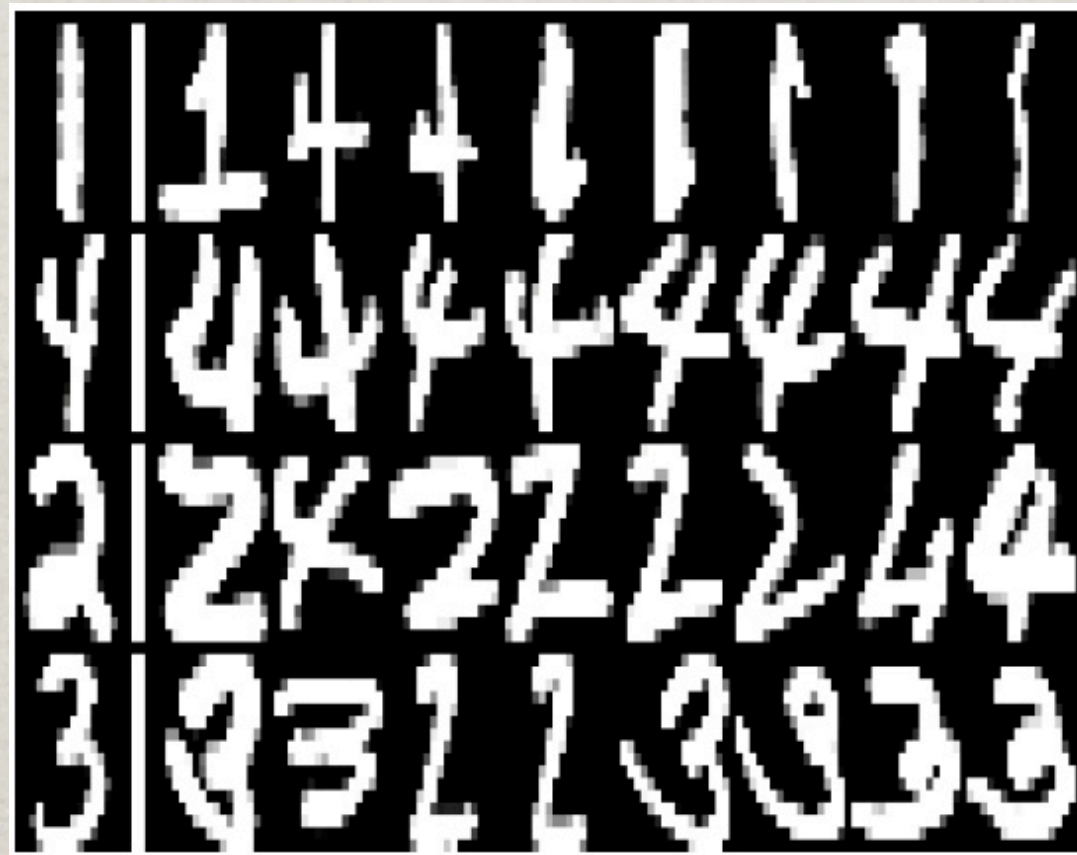
# Example: Handwritten digits

Database of 60,000 pictures, with  $28 \times 28$  pixels, of handwritten digits collected by USPS. Point cloud of 60,000 points in  $\mathbb{R}^{728}$ .



# Example: Handwritten digits

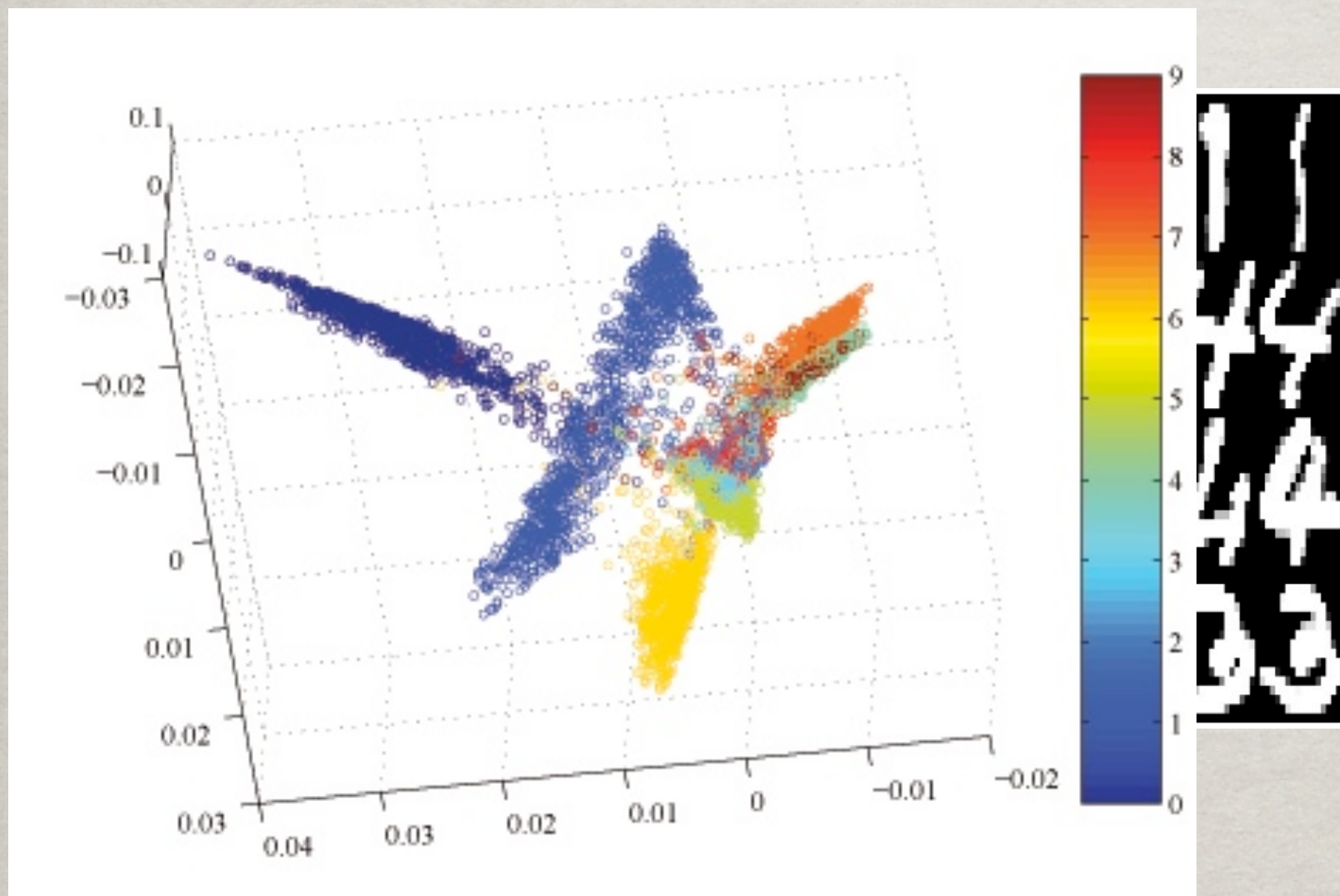
Database of 60,000 pictures, with  $28 \times 28$  pixels, of handwritten digits collected by USPS. Point cloud of 60,000 points in  $\mathbb{R}^{728}$ .





# Example: Handwritten digits

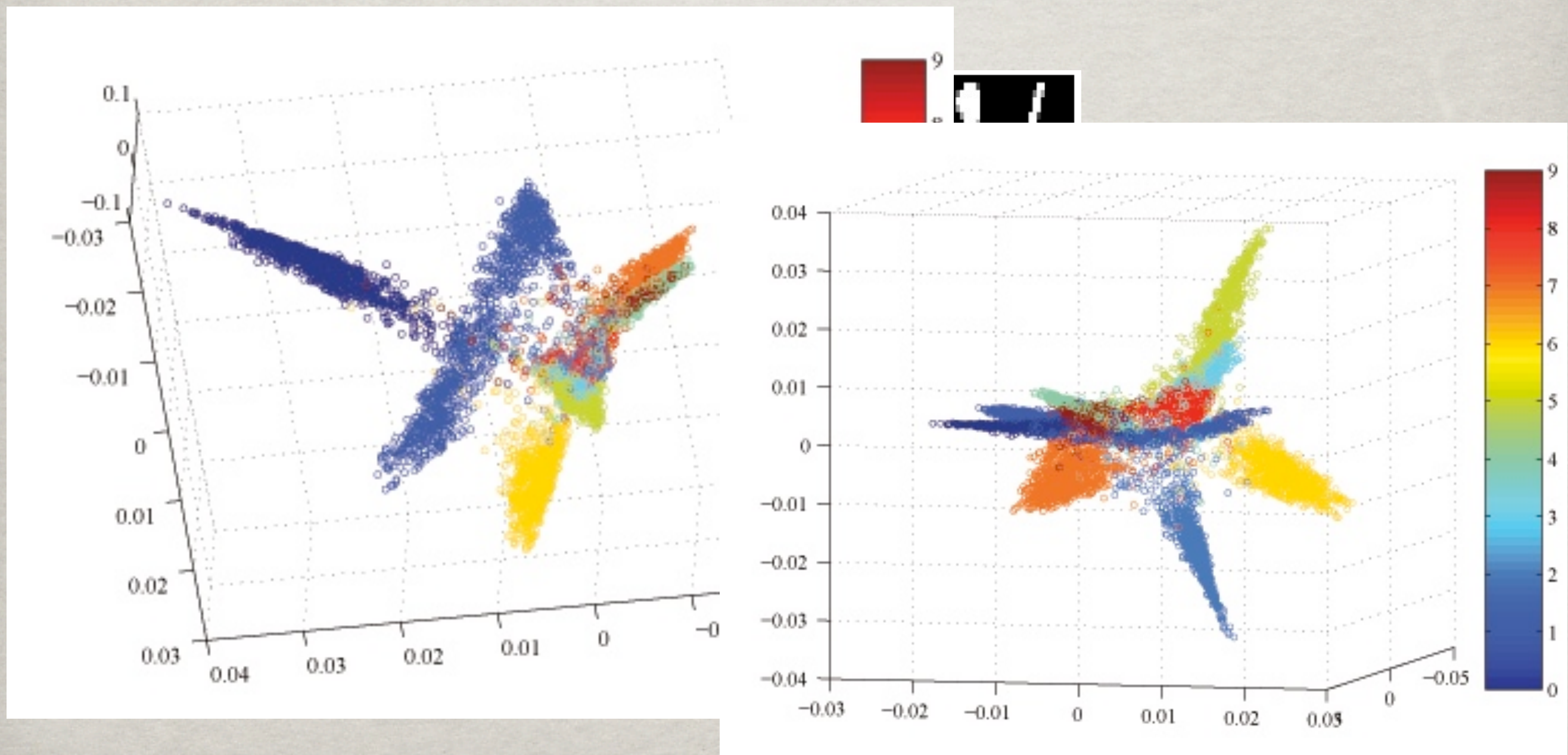
Database of 60,000 pictures, with  $28 \times 28$  pixels, of handwritten digits collected by USPS. Point cloud of 60,000 points in  $\mathbb{R}^{728}$ .





# Example: Handwritten digits

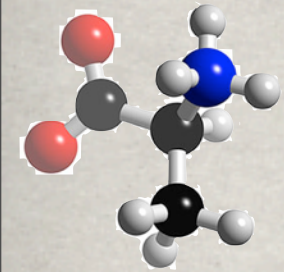
Database of 60,000 pictures, with  $28 \times 28$  pixels, of handwritten digits collected by USPS. Point cloud of 60,000 points in  $\mathbb{R}^{728}$ .





# Molecular Dynamics & F-P. equation

R.R.Coifman, I.G.Kevrekidis, S.Lafon, MM, B.Nadler, *Multiscale Model. Simul.*



Fokker-Planck equation & eigenfunctions

$$\frac{\partial p}{\partial t} = - \sum_i^{3N} \frac{\partial}{\partial x_i} \left( \frac{1}{\beta} \frac{\partial}{\partial x_i} + \frac{\partial E}{\partial x_i} \right) p = -\mathbf{H}_{\text{FP}} p$$

$\beta = 1/(k_B T)$ ,  $k_B$  is Boltzmann's constant

Under suitable conditions, it has discrete spectrum  $0 = \lambda_0 < \lambda_1 \leq \dots \lambda_k \ll \lambda_{k+1} \leq \dots$ , and fundamental solution with eigen-expansion

$$p_t(x, y) = \phi_0(x) + \sum_{j=1}^{+\infty} \psi_j(y) \phi_j(x) e^{-\lambda_j t}.$$

The dual system of eigenfunctions, which we pick as reaction coordinates, is

$$\psi_j(x) = \phi_j(x) / \phi_0(x).$$

With these normalizations,

$$d^{(t)}(x, y) = \|p_t(x, \cdot) - p_t(y, \cdot)\|_{L^2} = \sqrt{\sum_j e^{-\lambda_j t} |\psi_j(x) - \psi_j(y)|^2}$$



# Example: Molecular Dynamics Data

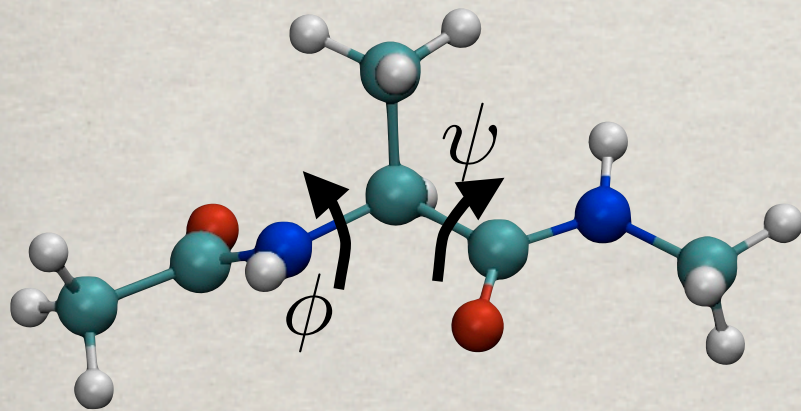
Joint with C. Clementi, M. Rohrdanz, W. Zheng

The dynamics of a small peptide (12 atoms with  $H$ -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 configurations is a point cloud in  $\mathbb{R}^{12 \times 3}$ .

$R^{36}$





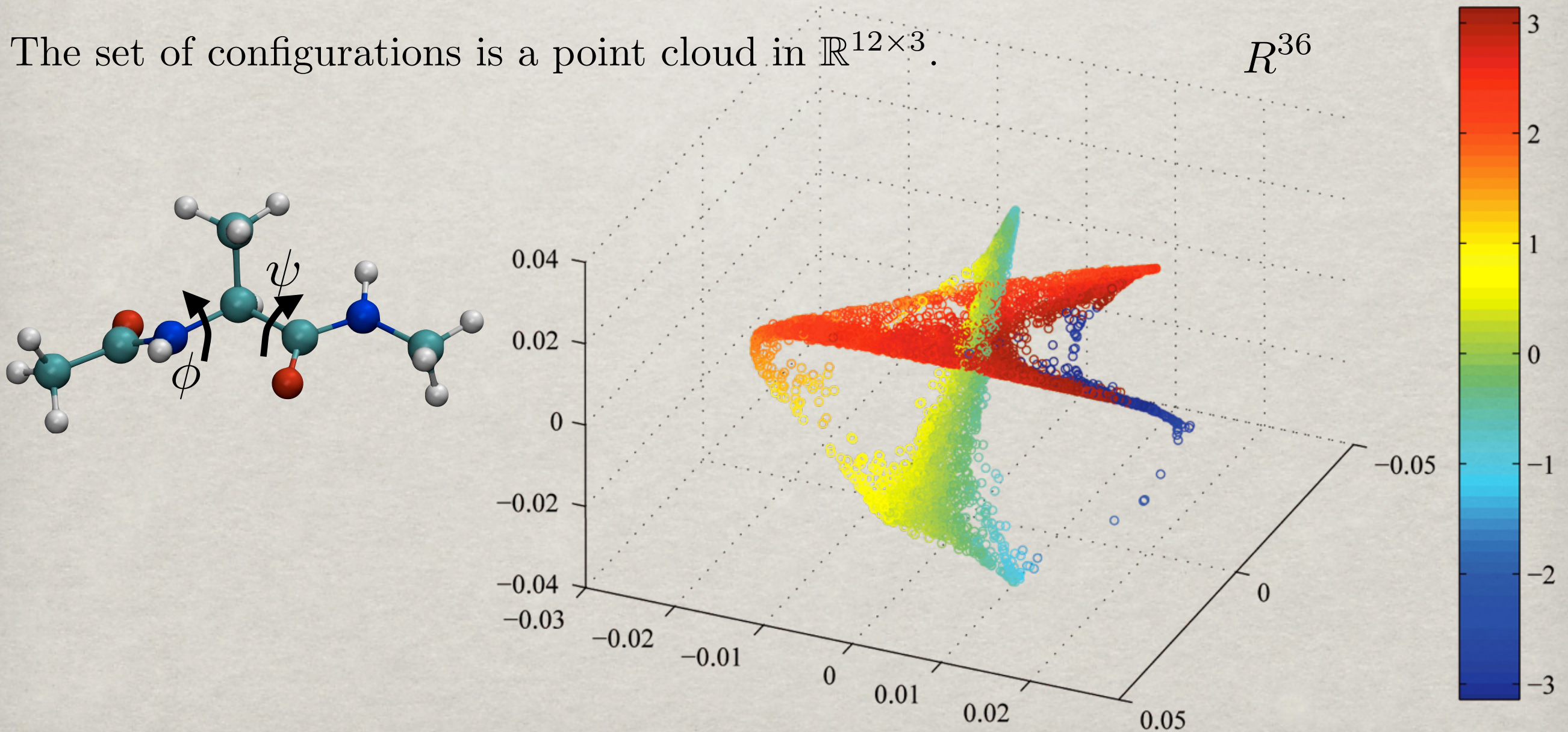
# Example: Molecular Dynamics Data

Joint with C. Clementi, M. Rohrdanz, W. Zheng

The dynamics of a small peptide (12 atoms with  $H$ -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 configurations is a point cloud in  $\mathbb{R}^{12 \times 3}$ .





# Intrinsic Dimension, Curvature, Local Scales

With A.V. Little, 2010

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  $\eta_i \sim_{\text{i.i.d.}} N$  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...
- . We will not only learn the intrinsic dimensionality, but also about the “natural scales” in the data: “immediate” applications to local scale selection, multiple plane models, dictionary learning...
- . Existing work somewhat unsatisfactory, both in theory and practice



# Existing approaches

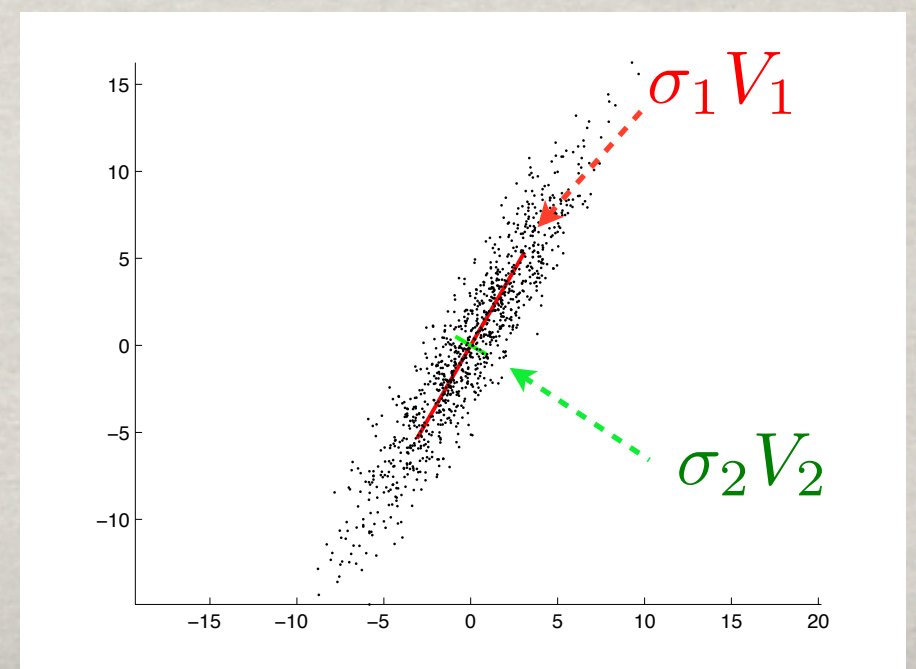
(high-level overview)

**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. Problematic choice for range of  $r$ . [Levina-Bickel; Haro-Randall-Sapiro; Carter-Hero (x3); Costa-Hero; Camastra-Vinciarelli; Cao-Haralick; Raginsky-Lazebnik; Takens; Hein-Audibert; Bruske-Sommer...] Sample complexity:  $n \sim 2^k$ .

**Principal Component Analysis:** if  $X_n$  is the  $n \times D$  matrix with the samples, let  $\text{cov}(X_n) = \frac{1}{n} X_n^T X_n = \frac{1}{n} V \Sigma^2 V^T$  with the diagonal  $\Sigma = \text{diag}(\sigma_1, \dots, \sigma_D)$ . The plane  $\pi_i$  spanned by the top  $i$   $V_m$ 's minimizes

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

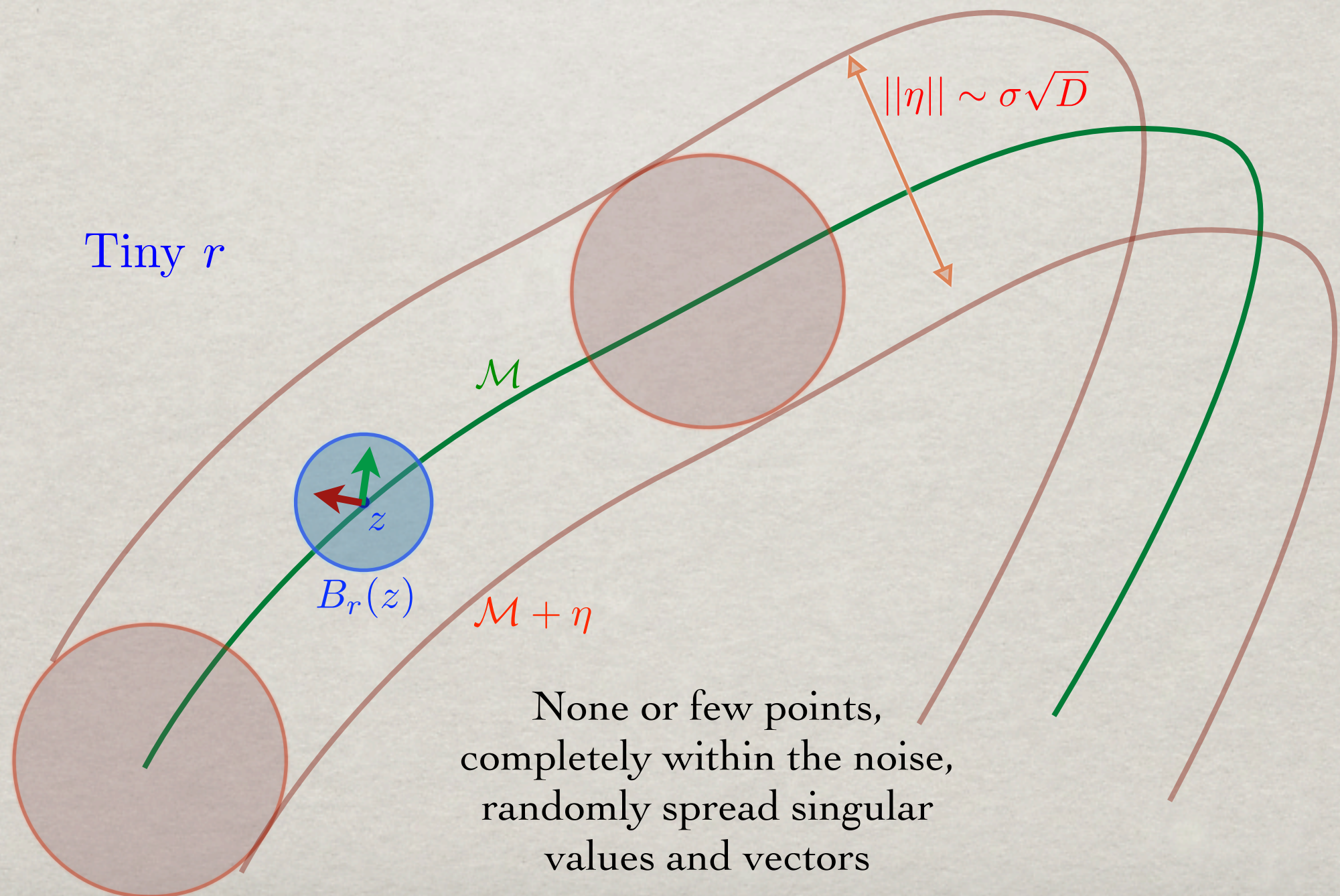
Great if data sampled from a linear subspace. If sampled from manifold, do this for data in a “small” ball of radius  $r$  [Fukunaga '67]. Problematic choice for  $r$ . Multiple  $r$ 's? [Kirby] Sample complexity: need  $n \gtrsim k \log k$  for estimating a rank  $k$  covariance.





# Multiscale SVD

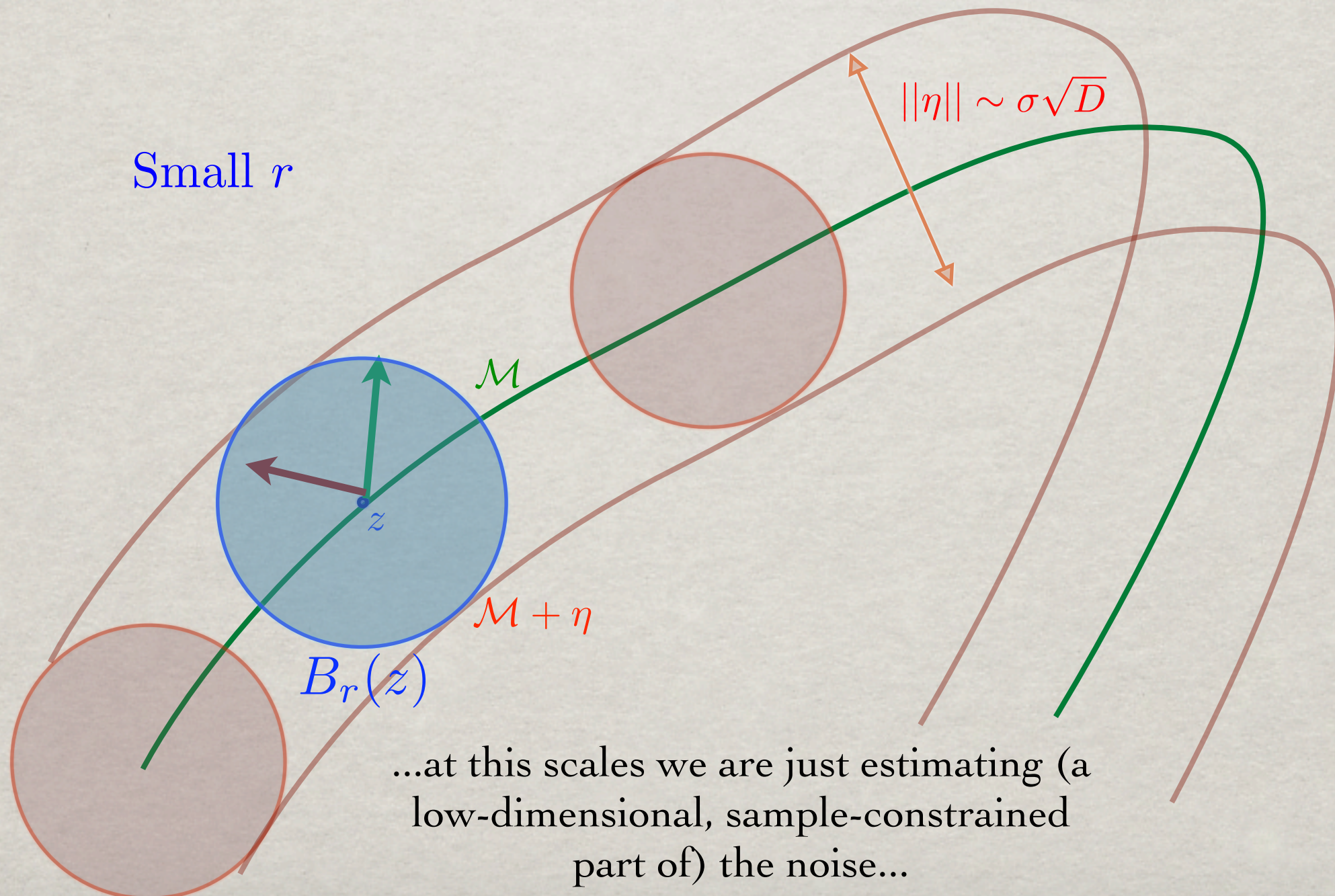
Let  $X_n$  be  $n$  points sampled from  $\mathcal{M}$ , corrupted by  $\eta \sim \sigma\mathcal{N}(0, I_D)$ . Note that  $\|\eta\| \sim \sigma\sqrt{D}$ . Let  $\sigma_i^{z,r}$  be the  $i$ -th singular value of the covariance matrix of  $X_n$  restricted to  $B_z(r)$ .





# Multiscale SVD

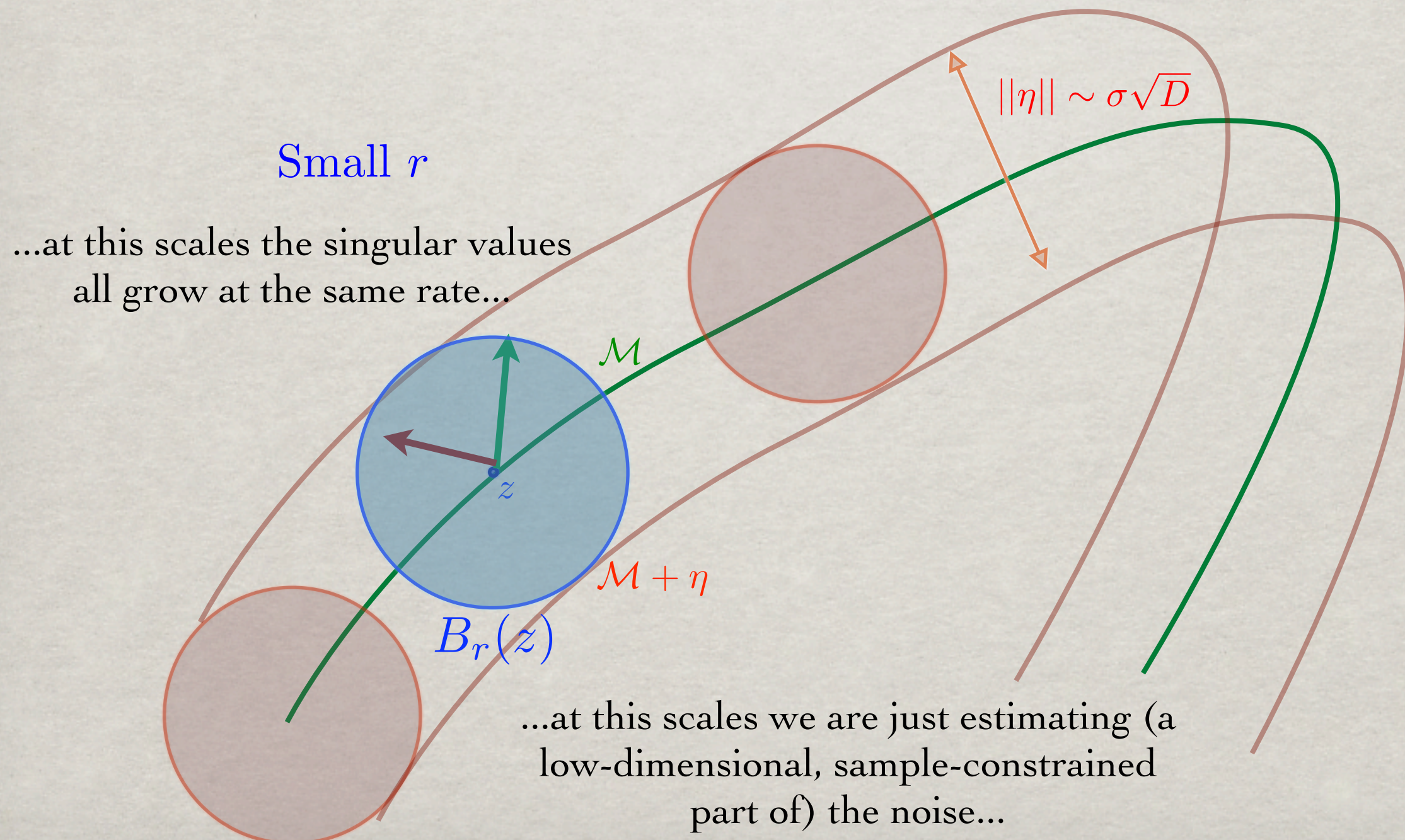
Let  $X_n$  be  $n$  points sampled from  $\mathcal{M}$ , corrupted by  $\eta \sim \sigma\mathcal{N}(0, I_D)$ . Note that  $\|\eta\| \sim \sigma\sqrt{D}$ . Let  $\sigma_i^{z,r}$  be the  $i$ -th singular value of the covariance matrix of  $X_n$  restricted to  $B_z(r)$ .





# Multiscale SVD

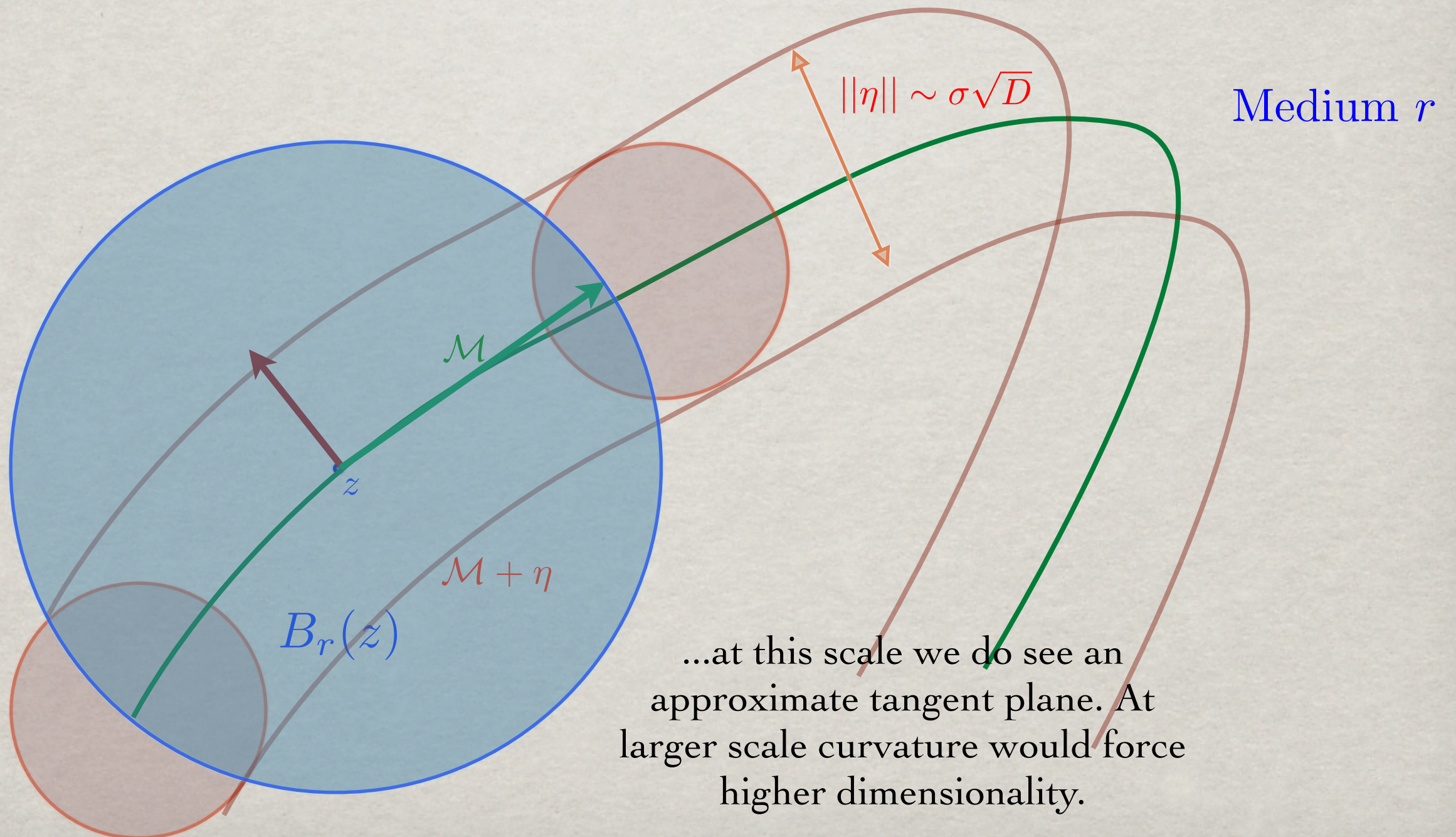
Let  $X_n$  be  $n$  points sampled from  $\mathcal{M}$ , corrupted by  $\eta \sim \sigma\mathcal{N}(0, I_D)$ . Note that  $\|\eta\| \sim \sigma\sqrt{D}$ . Let  $\sigma_i^{z,r}$  be the  $i$ -th singular value of the covariance matrix of  $X_n$  restricted to  $B_z(r)$ .





# Multiscale SVD

Let  $X_n$  be  $n$  points sampled from  $\mathcal{M}$ , corrupted by  $\eta \sim \sigma\mathcal{N}(0, I_D)$ . Note that  $\|\eta\| \sim \sigma\sqrt{D}$ . Let  $\sigma_i^{z,r}$  be the  $i$ -th singular value of the covariance matrix of  $X_n$  restricted to  $B_z(r)$ .

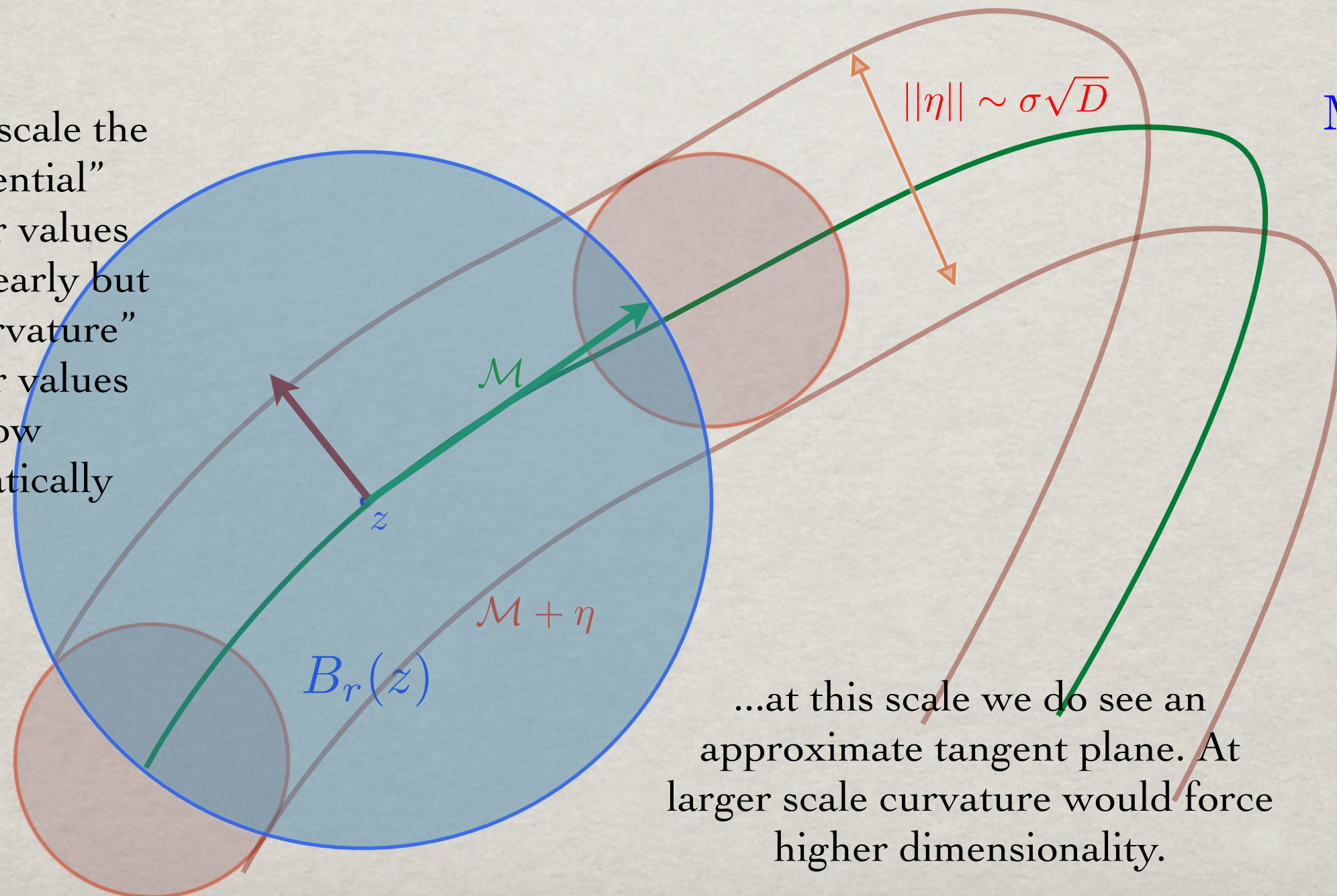




# Multiscale SVD

Let  $X_n$  be  $n$  points sampled from  $\mathcal{M}$ , corrupted by  $\eta \sim \sigma\mathcal{N}(0, I_D)$ . Note that  $\|\eta\| \sim \sigma\sqrt{D}$ . Let  $\sigma_i^{z,r}$  be the  $i$ -th singular value of the covariance matrix of  $X_n$  restricted to  $B_z(r)$ .

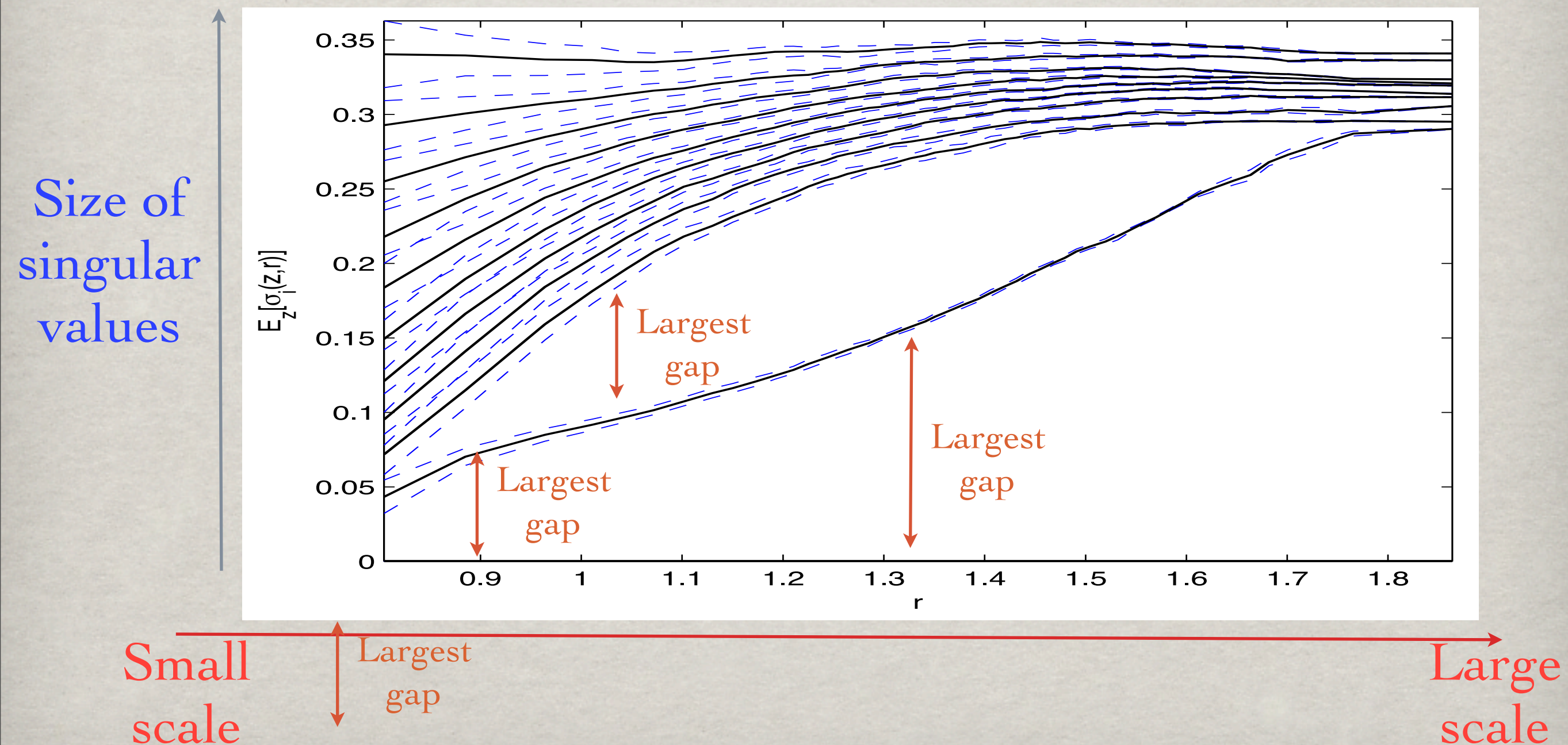
...at this scale the  
“tangential”  
singular values  
grow linearly but  
the “curvature”  
singular values  
grow  
quadratically





# Multiscale SVD: Sphere

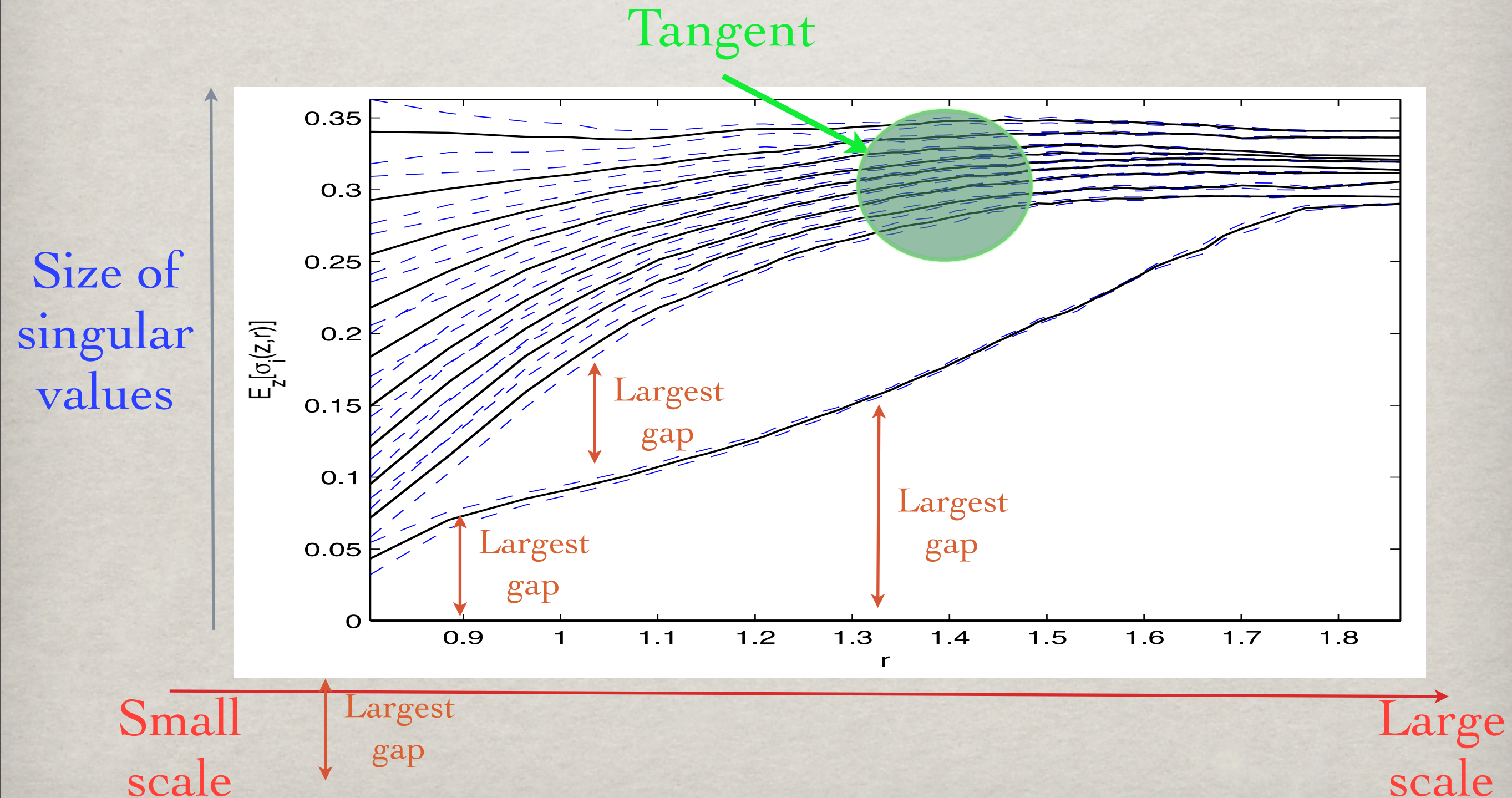
Example: consider  $\mathbb{S}^9(100, 1000, 0)$ : 1000 points uniformly samples on a 9-dimensional unit sphere, embedded in 100 dimensions, with no noise.





# Multiscale SVD: Sphere

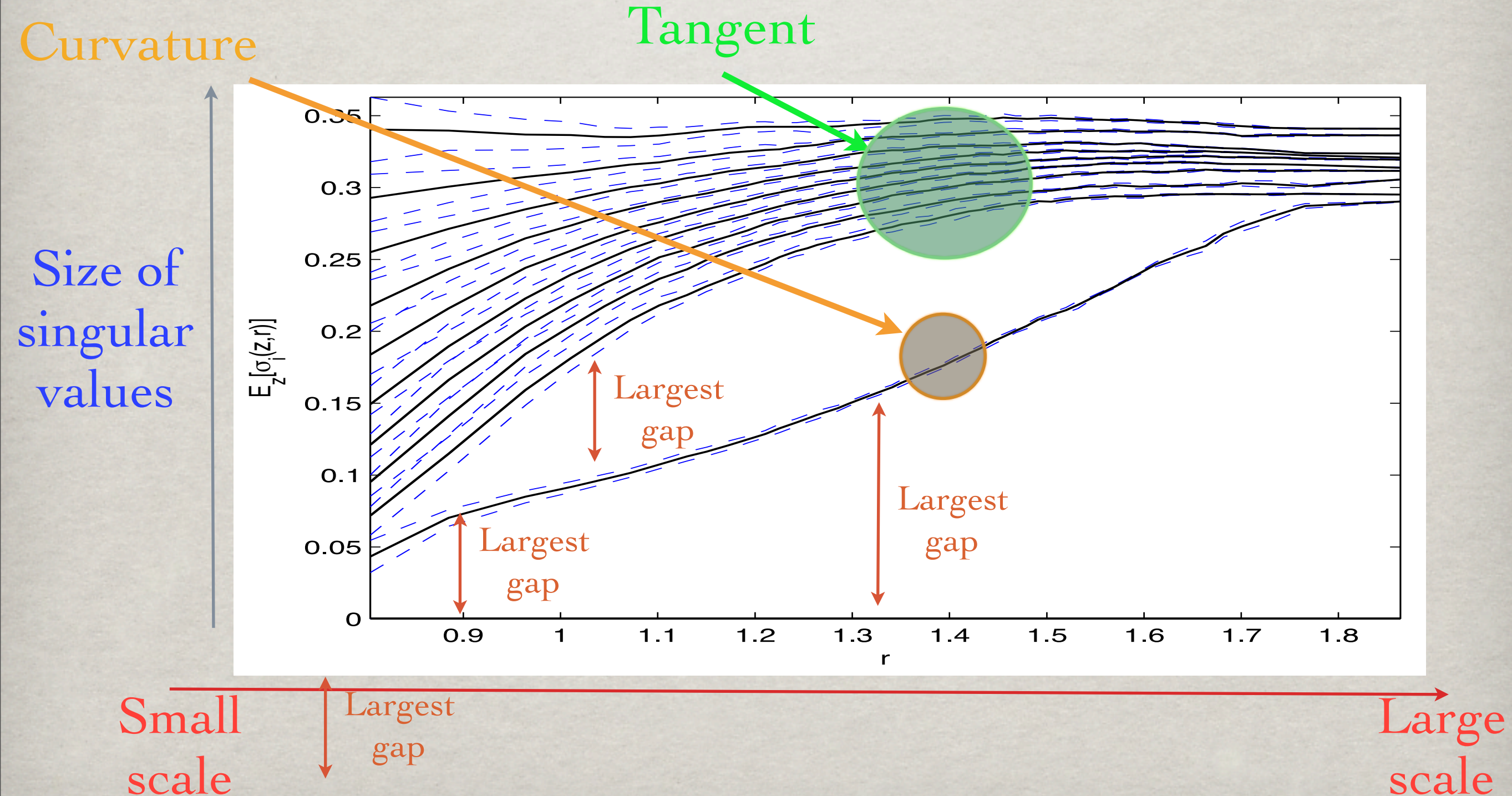
Example: consider  $\mathbb{S}^9(100, 1000, 0)$ : 1000 points uniformly samples on a 9-dimensional unit sphere, embedded in 100 dimensions, with no noise.





# Multiscale SVD: Sphere

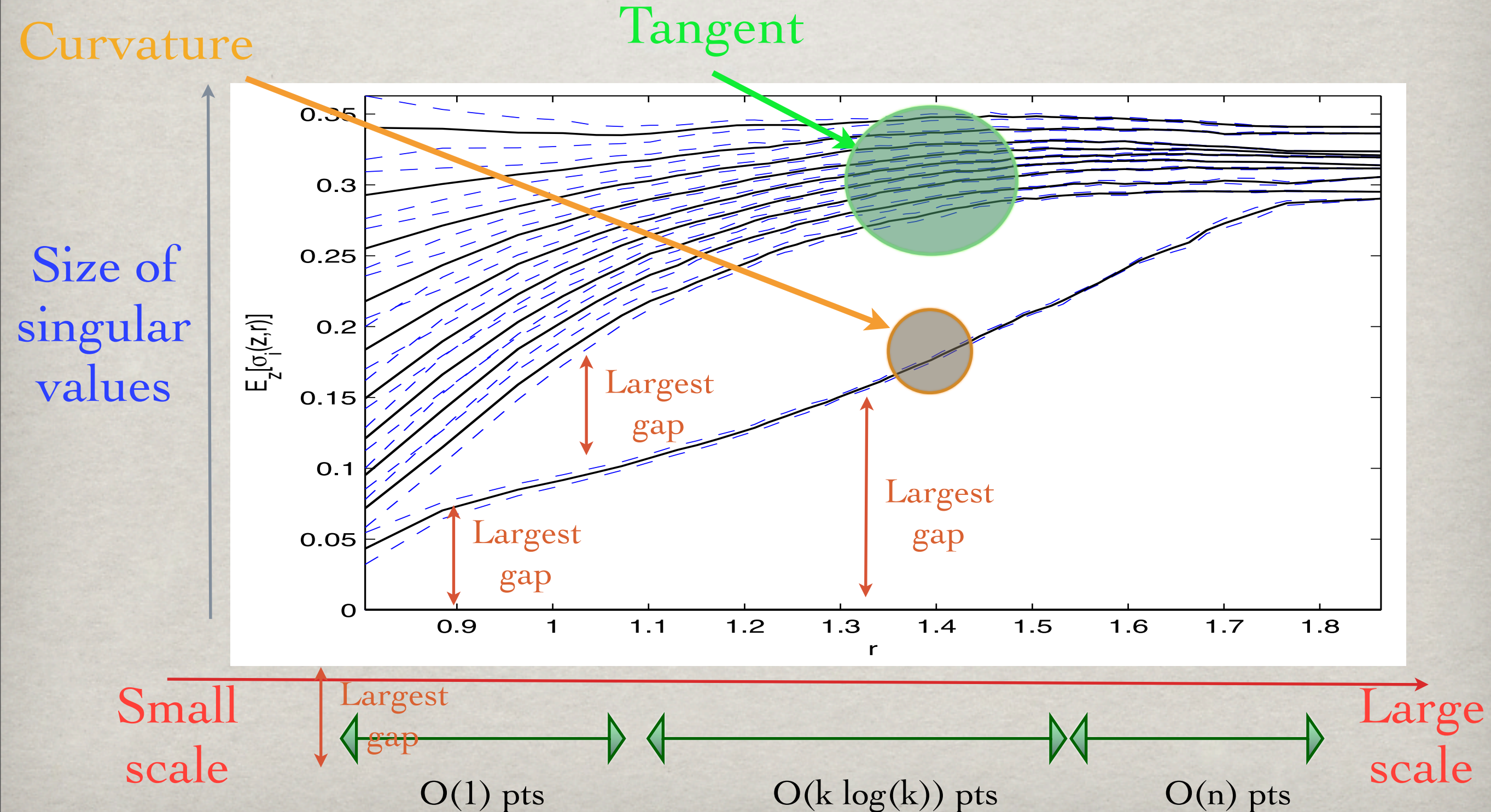
Example: consider  $\mathbb{S}^9(100, 1000, 0)$ : 1000 points uniformly samples on a 9-dimensional unit sphere, embedded in 100 dimensions, with no noise.





# Multiscale SVD: Sphere

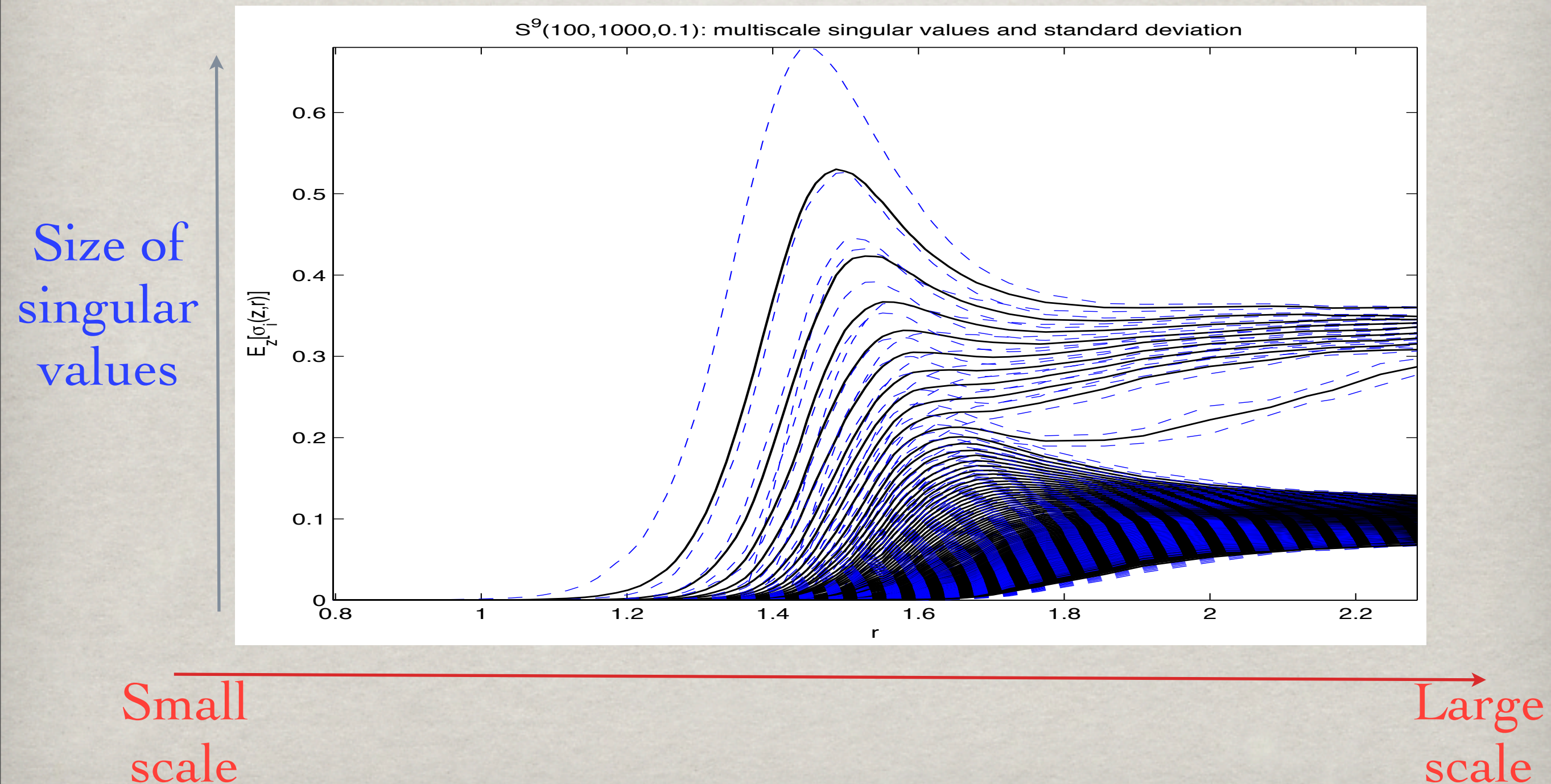
Example: consider  $\mathbb{S}^9(100, 1000, 0)$ : 1000 points uniformly samples on a 9-dimensional unit sphere, embedded in 100 dimensions, with no noise.





# Multiscale SVD: Sphere+noise

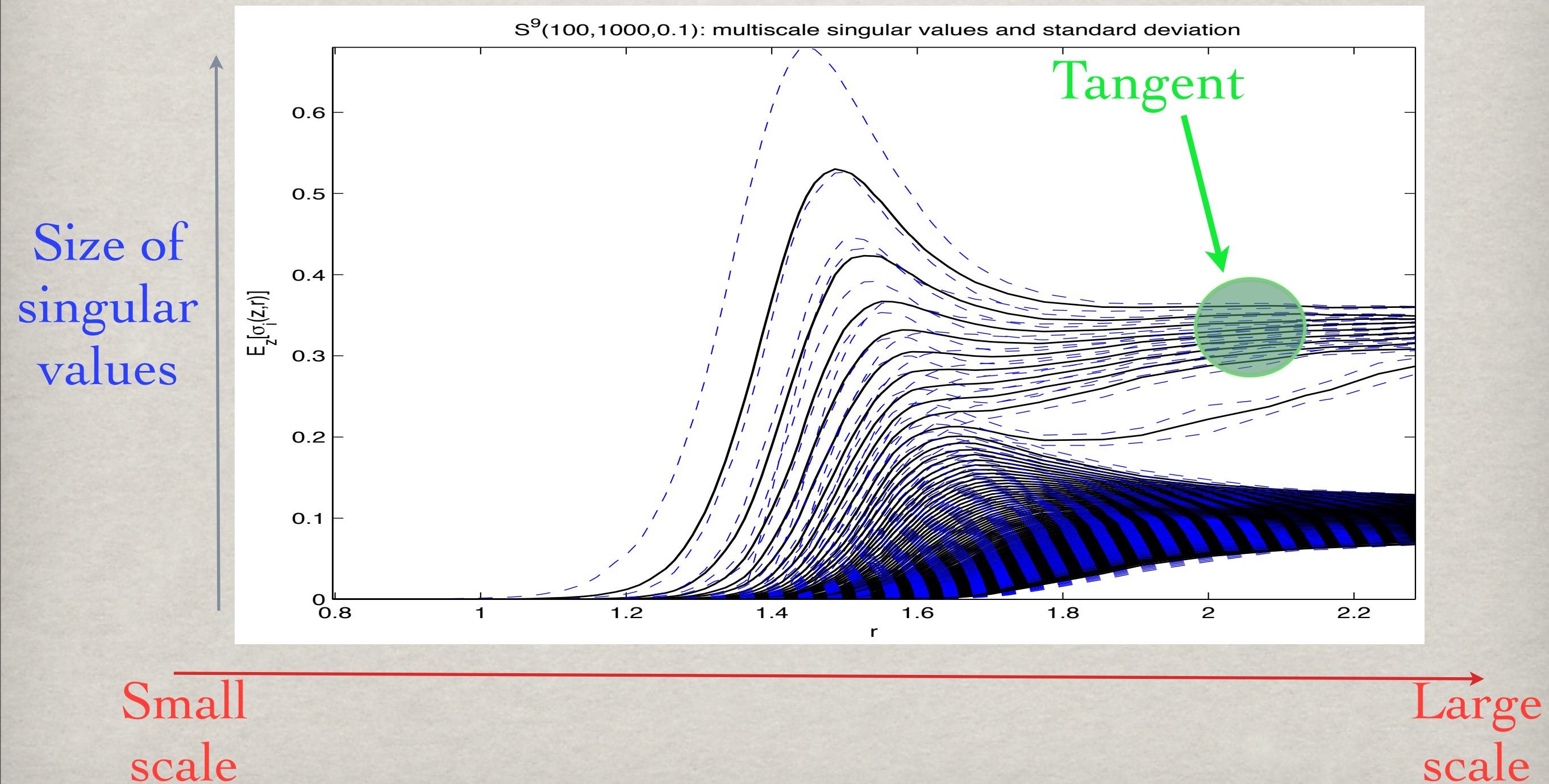
Example: consider  $S^9(100, 1000, 0.1)$ : 1000 points uniformly samples on a 9-dimensional unit sphere, embedded in 100 dimensions, with Gaussian noise  $\mathcal{N}(0, 0.1I_{100})$ . Observe that  $\mathbb{E}[\|\eta\|^2] \sim 0.1^2 \cdot 100 = 1$ .





# Multiscale SVD: Sphere+noise

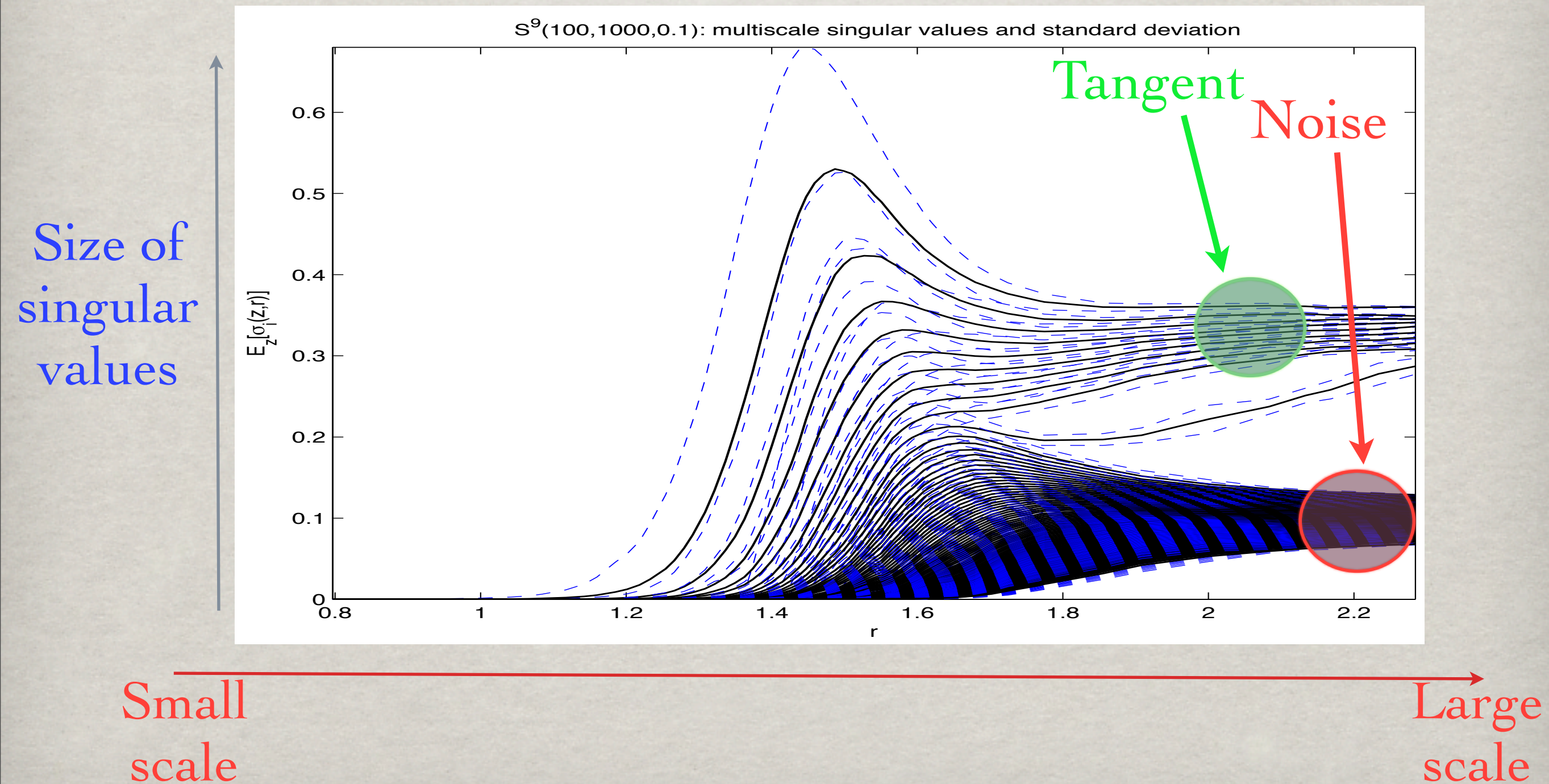
Example: consider  $S^9(100, 1000, 0.1)$ : 1000 points uniformly samples on a 9-dimensional unit sphere, embedded in 100 dimensions, with Gaussian noise  $\mathcal{N}(0, 0.1I_{100})$ . Observe that  $\mathbb{E}[\|\eta\|^2] \sim 0.1^2 \cdot 100 = 1$ .





# Multiscale SVD: Sphere+noise

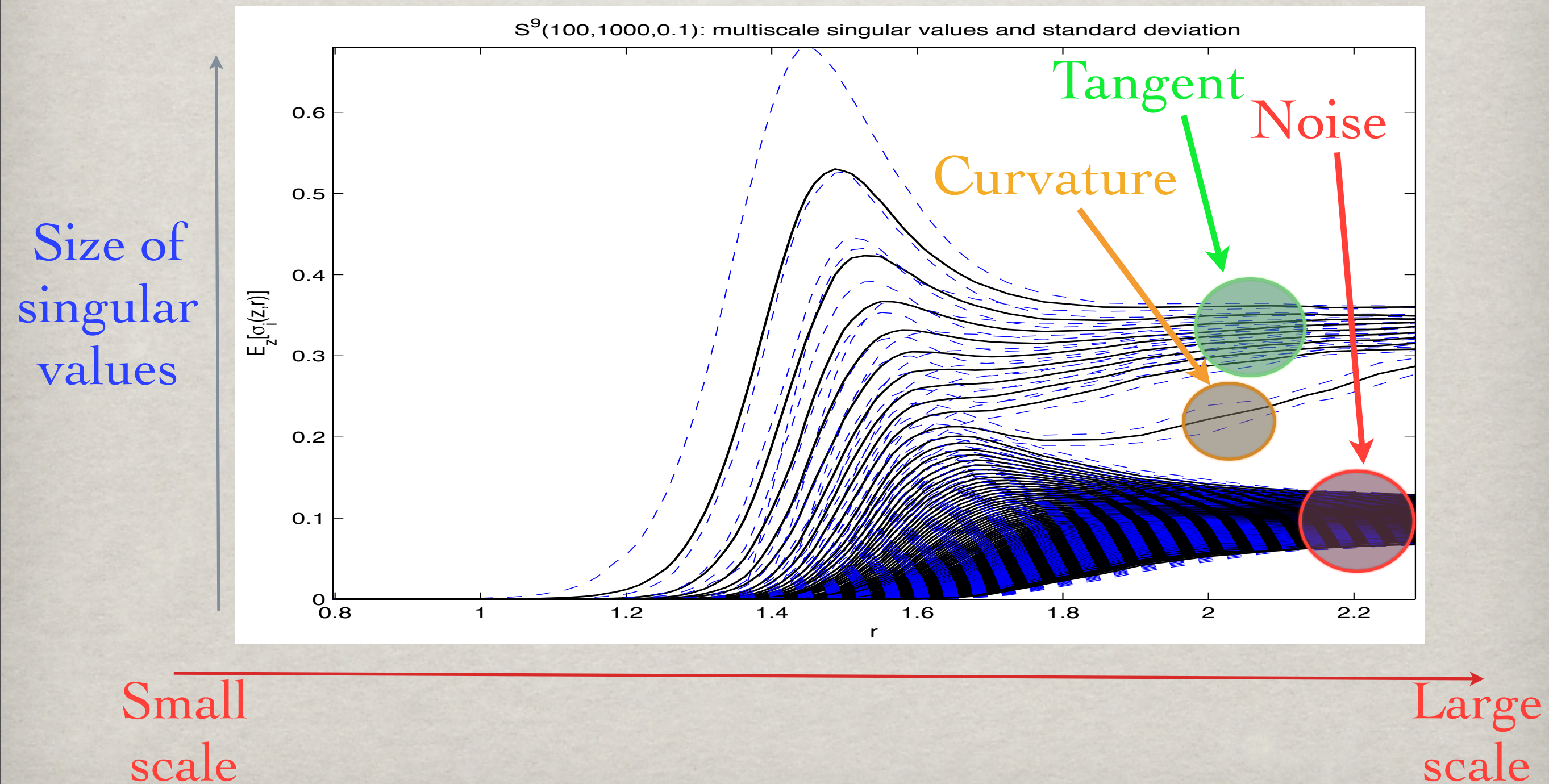
Example: consider  $S^9(100, 1000, 0.1)$ : 1000 points uniformly samples on a 9-dimensional unit sphere, embedded in 100 dimensions, with Gaussian noise  $\mathcal{N}(0, 0.1I_{100})$ . Observe that  $\mathbb{E}[\|\eta\|^2] \sim 0.1^2 \cdot 100 = 1$ .





# Multiscale SVD: Sphere+noise

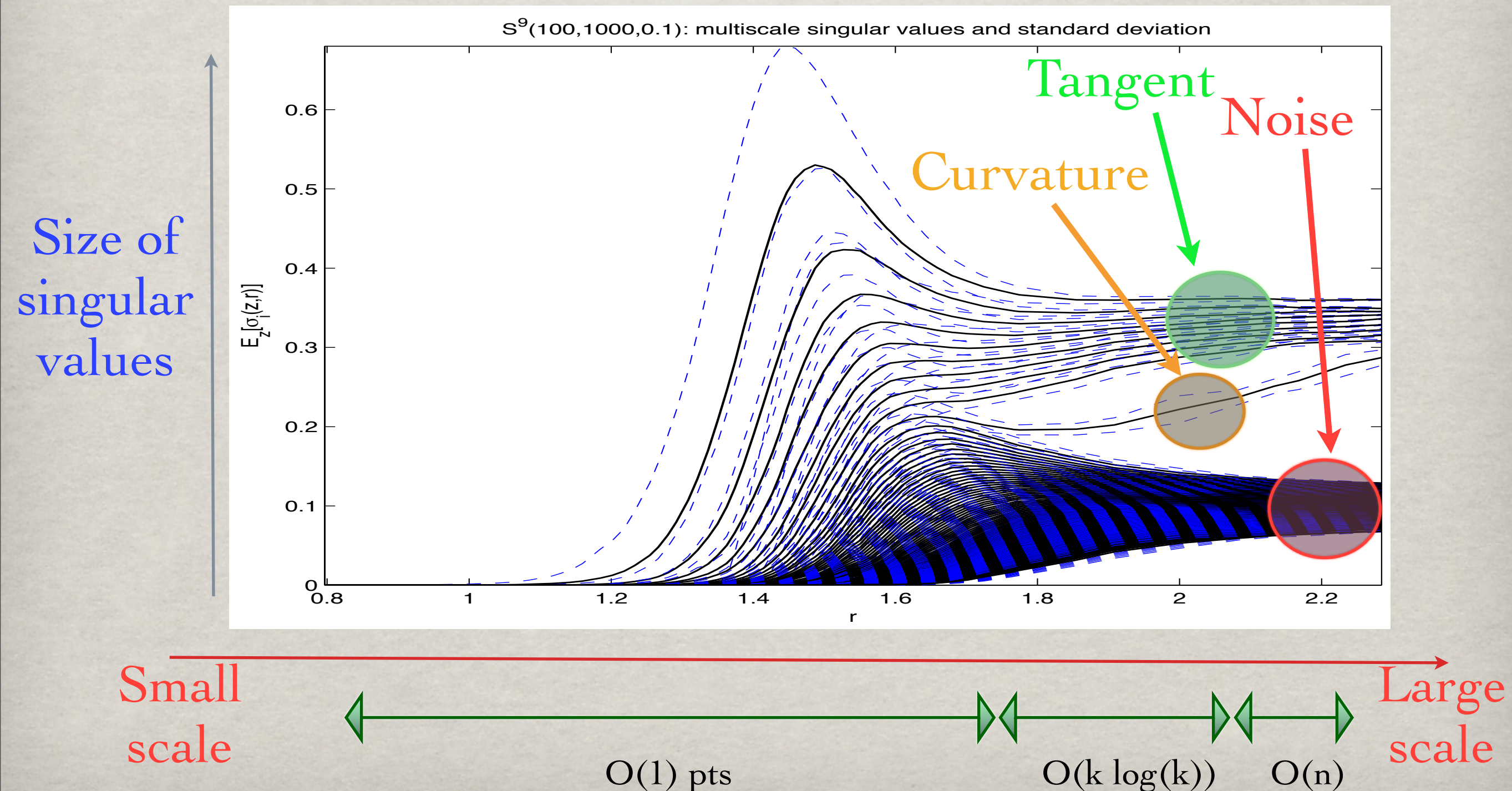
Example: consider  $S^9(100, 1000, 0.1)$ : 1000 points uniformly samples on a 9-dimensional unit sphere, embedded in 100 dimensions, with Gaussian noise  $\mathcal{N}(0, 0.1I_{100})$ . Observe that  $\mathbb{E}[\|\eta\|^2] \sim 0.1^2 \cdot 100 = 1$ .





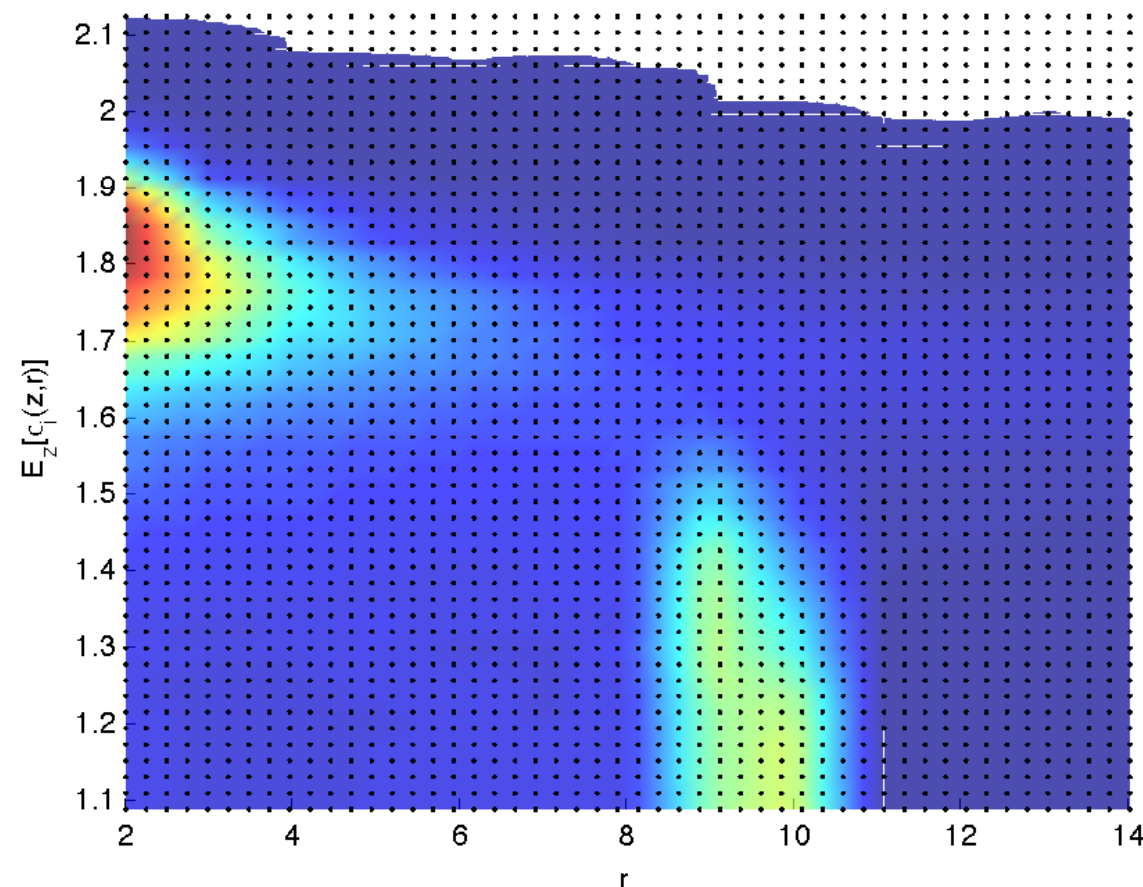
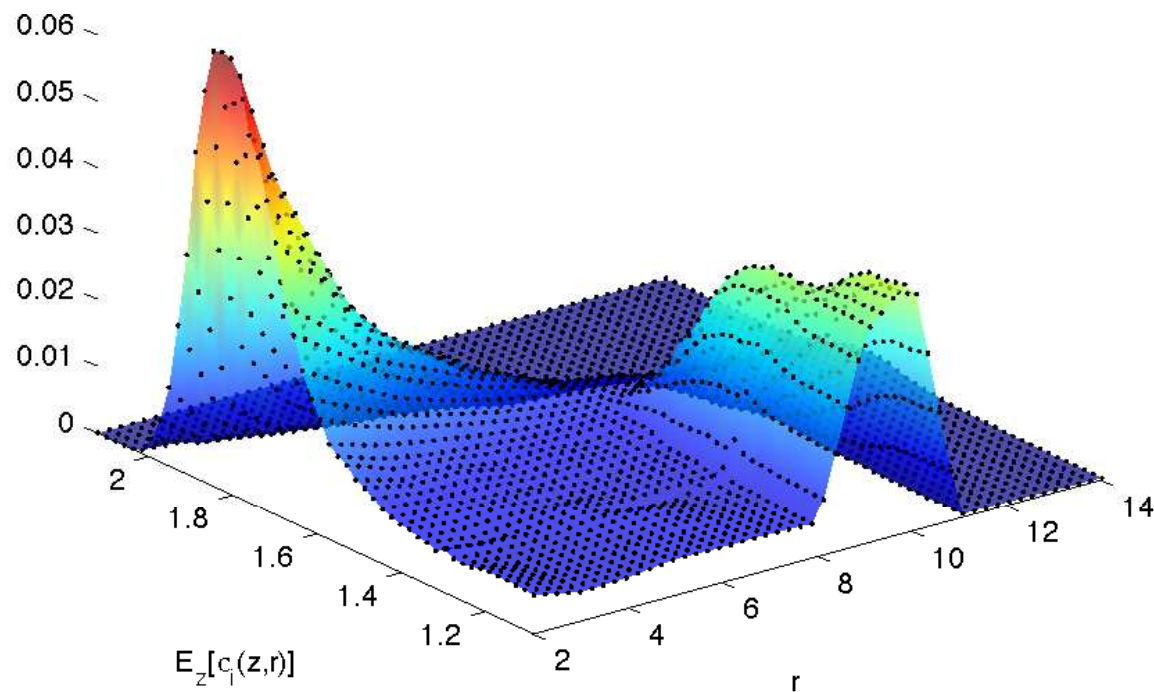
# Multiscale SVD: Sphere+noise

Example: consider  $S^9(100, 1000, 0.1)$ : 1000 points uniformly samples on a 9-dimensional unit sphere, embedded in 100 dimensions, with Gaussian noise  $\mathcal{N}(0, 0.1I_{100})$ . Observe that  $\mathbb{E}[\|\eta\|^2] \sim 0.1^2 \cdot 100 = 1$ .

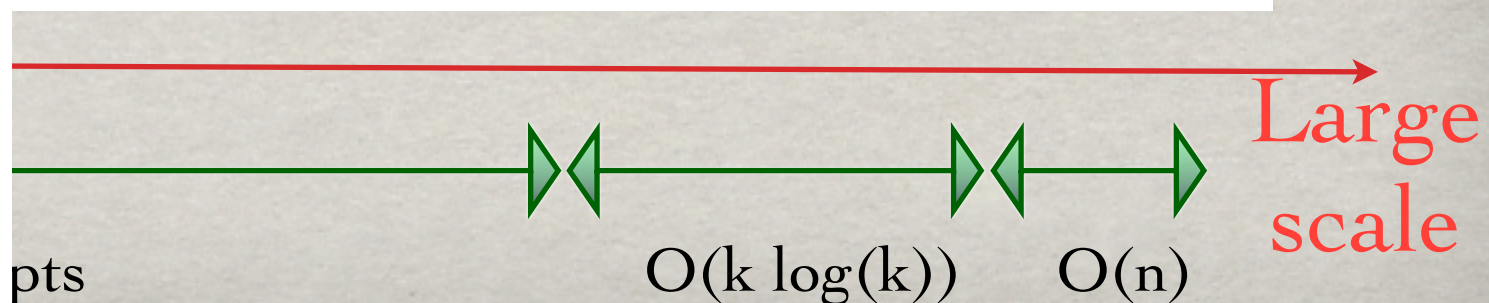
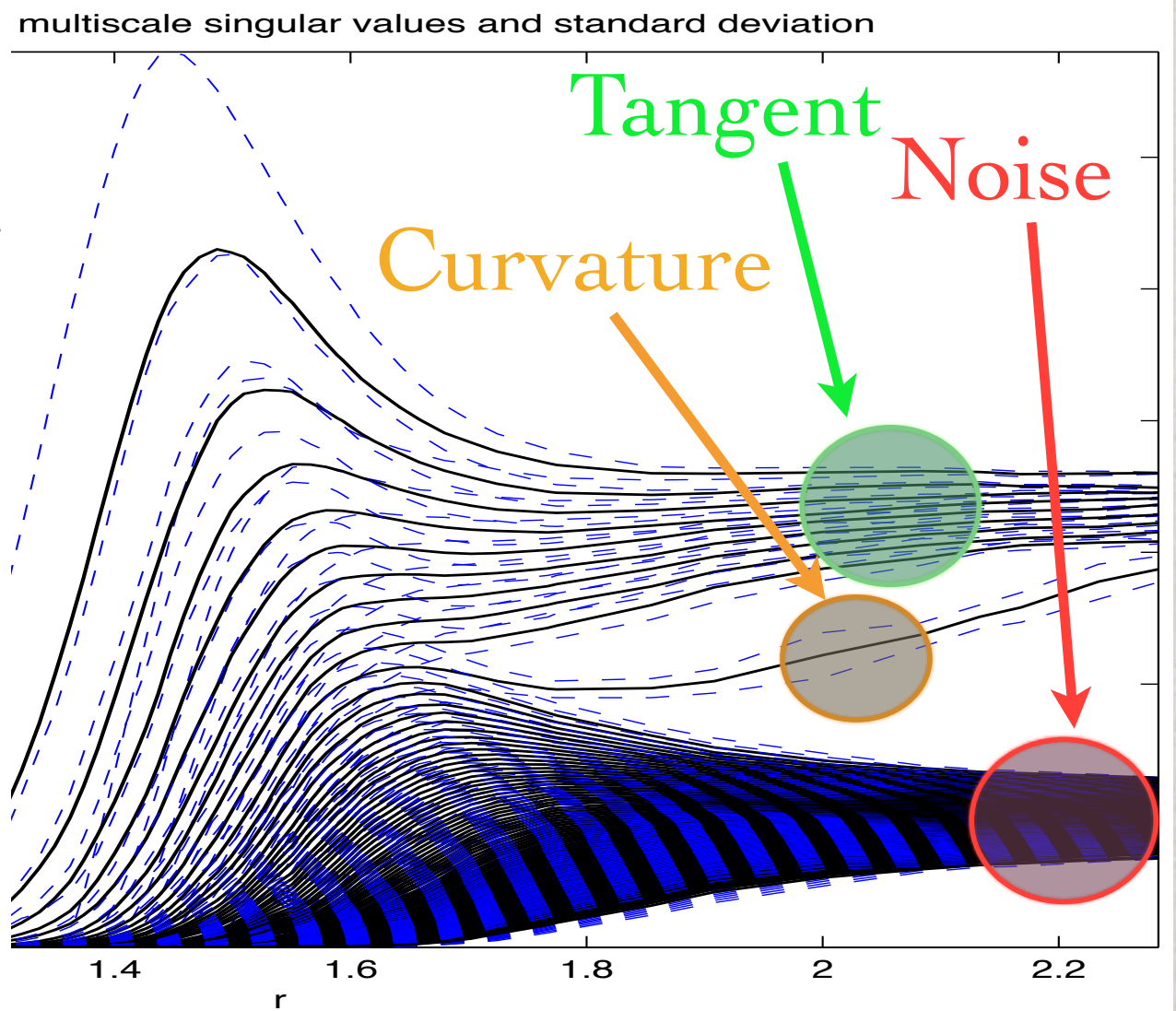




# Multiscale SVD: Sphere+noise



l): 1000 points uniformly samples on a 9-  
d in 100 dimensions, with Gaussian noise  
 $\sim 0.1^2 \cdot 100 = 1$ .





# Sketch of results

Some volume-based algorithms are **consistent** ( $n \rightarrow +\infty$ ). Random matrix theory algorithms typically succeed in the setting  $n, D \rightarrow +\infty$ , with  $\frac{n}{D} \rightarrow \gamma$ . We seek **finite sample results, with high probability**: given  $n, D$  and other parameters of the problem, promise that with probability at least  $1 - e^{-ct^2}$  we return the correct answer.

We prove that: if  $(R_{\min}, R_{\max})$  is the range of scales for which the “manifold” looks “flat” in  $k$  directions and “thin” in the others, provided that

$\mathcal{M}$  has small “curvature”  $\kappa$  + the noise  $\eta$  has small std  $\sigma$ ,

$\Rightarrow$  w.h.p. as soon as  $n_r \gtrsim_t k \log k$  (for  $r$  a good scale), a slightly smaller range of scales survives sampling and noise.

Moreover, if  $\kappa \sim 1$  and  $\sigma \sim D^{-\frac{1}{2}}$ , then our results are **dimension-free**, i.e. the above is true w.h.p. for  $n \gtrsim k \log k$  with the implicit constants independent of  $k, D$ .

**MUCH MORE GENERAL THAN MANIFOLDS**  
**ALLOWS FOR DIFFERENT DIMENSIONS AT DIFFERENT POINTS/SCALES**

The proof uses combination of spectral theory for random matrices and covariance matrix estimation, adapted to analyze this “low-rank” + small perturbation point clouds...

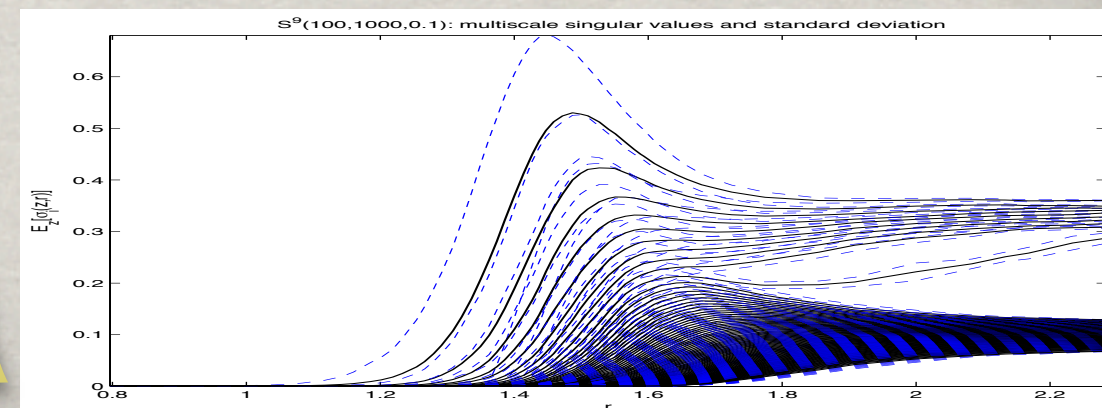


# Algorithm

- (1) **compute**  $\sigma_i^{(z,r)}$ , for each  $z \in \mathcal{M}$ ,  $r > 0$ ,  $i = 1, \dots, D$ .
- (2) **estimate the noise size**  $\sigma$ , obtained from the bottom S.V.'s which do not grow with  $r$ . Split the S.V.'s into noise S.V.'s and non-noise S.V.'s.
- (3) **identify a range of scales** where the noise S.V.'s are small compared to the other S.V.'s.
- (4) **estimate**, in the range of scales identified, which S.V.'s, among the non-noise S.V.'s, correspond to tangent directions and which ones correspond to curvatures, by comparing the growth rate as being linear or quadratic in  $r^2$ .

**Computational considerations:** by constructing multiscale nets, instead of computing at all scales and all locations, and randomized SVD [Martinsson, Rokhlin, Tropp, Tygert], the cost of the algorithm, assuming that finding nearest neighbors is (after preprocessing)  $O(\log n)$ , becomes  $O(KDn \log n)$ , where  $K$  is a given upper bound on  $k$ .

Extensive experiments show that this is the case  
Very competitive in terms of speed against many other algorithms  
Scales very well with  $n$  as well as  $K$





# Comparisons: existing algorithms

Smoothing -	K. Carter, A. Hero, Variance reduction with neighborhood smoothing for local intrinsic dimension estimation, Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on (2008) 3917–3920.
RPMM -	G. Haro, G. Randall, G. Sapiro, Translated poisson mixture model for stratification learning, Int. J. Comput. Vision 80 (3) (2008) 358–374.
MLE -	E. Levina, P. J. Bickel, Maximum likelihood estimation of intrinsic dimension, in: L. K. Saul, Y. Weiss, L. Bottou (Eds.), Advances in Neural Information Processing Systems 17, MIT Press, Cambridge, MA, 2005, pp. 777–784.
Debias -	K. M. Carter, A. O. Hero, R. Raich, De-biasing for intrinsic dimension estimation, Statistical Signal Processing, 2007. SSP '07. IEEE/SP 14th Workshop on (2007) 601–605.
KNN -	J. Costa, A. Hero, Geodesic entropic graphs for dimension and entropy estimation in manifold learning, Signal Processing, IEEE Transactions on 52 (8) (2004) 2210–2221.
IDE -	M. Hein, Y. Audibert, Intrinsic dimensionality estimation of submanifolds in euclidean space, in: S. W. De Raedt, L. (Ed.), ICML Bonn, 2005, pp. 289 – 296.
MFA -	M. Chen, J. Silva, J. Paisley, C. Wang, D. Dunson, L. Carin, Compressive sensing on manifolds using a nonparametric mixture of factor analyzers: Algorithm and performance bounds, IEEE Trans. Signal Processing.

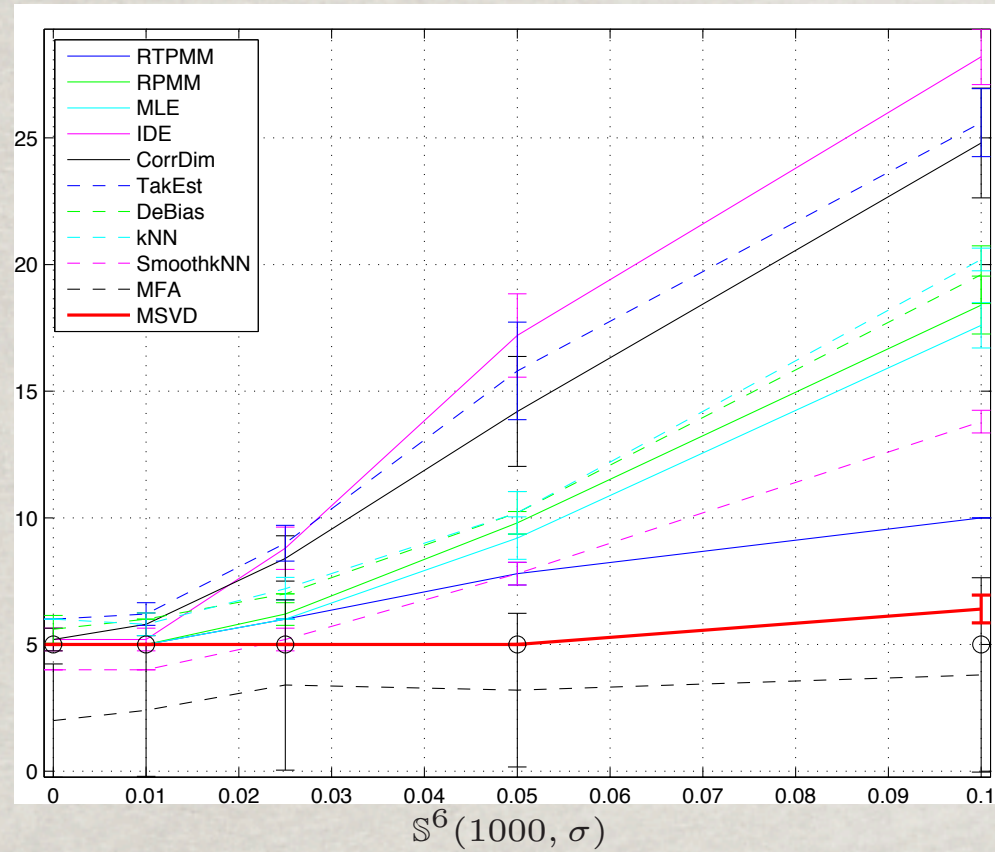
+ classical correlation dimension and Takens estimator

All the parameters in our algorithm are **fixed** in all the examples, comparisons, toy and real data sets, etc...

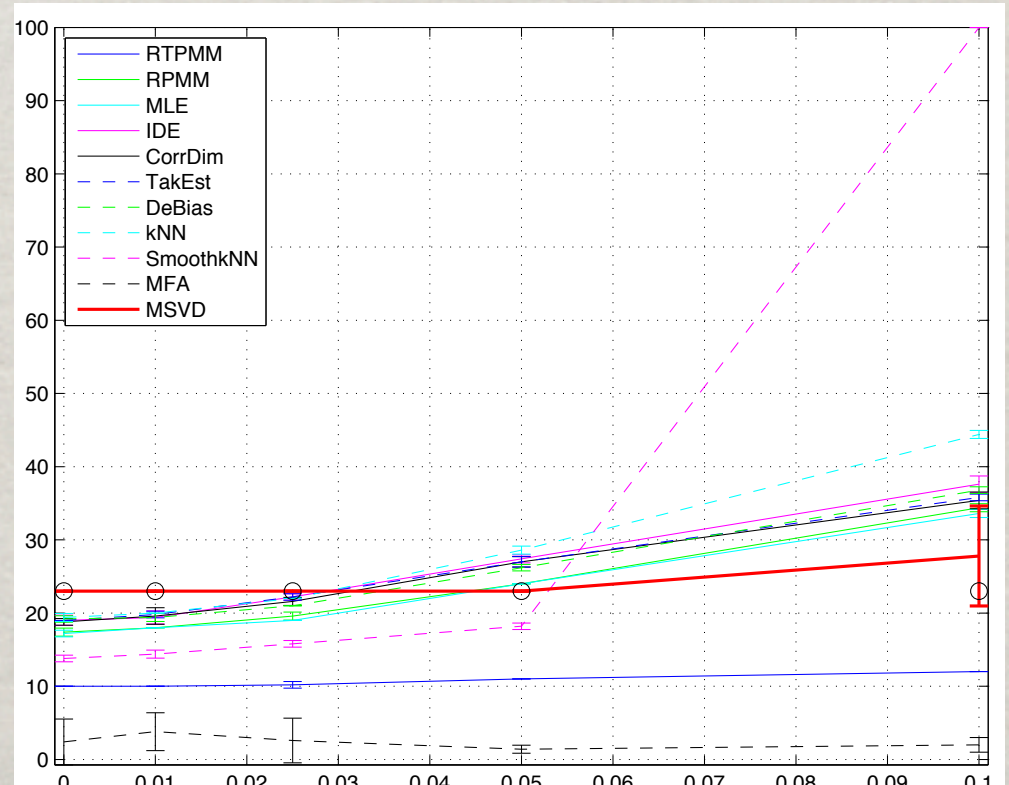
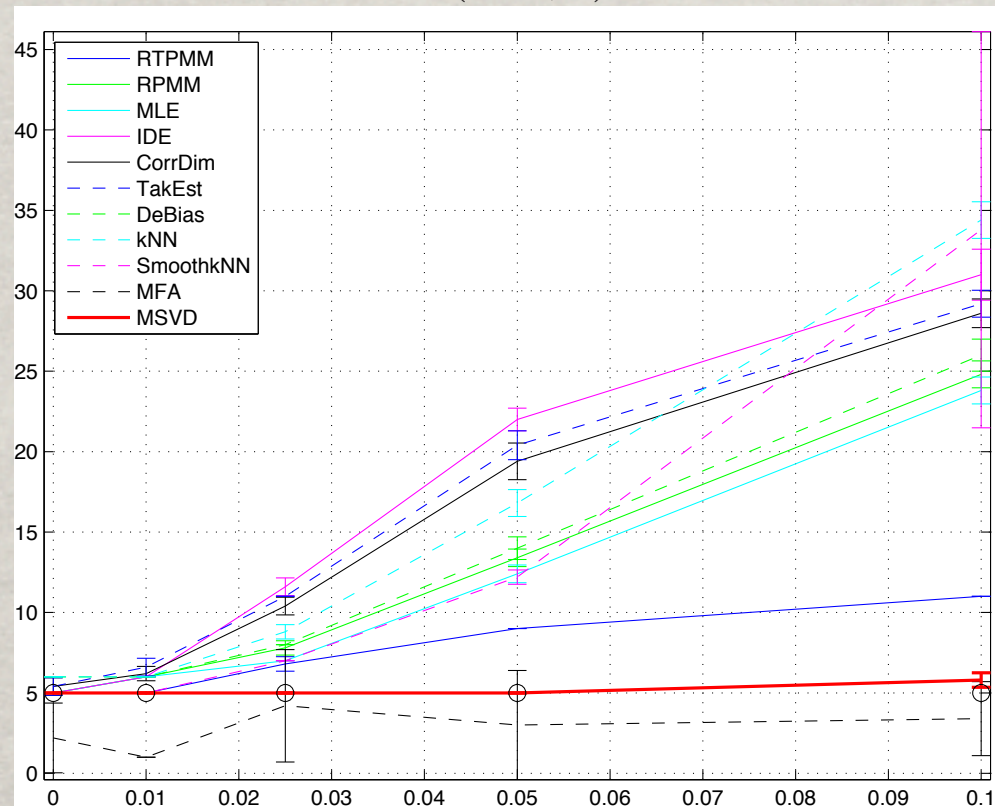
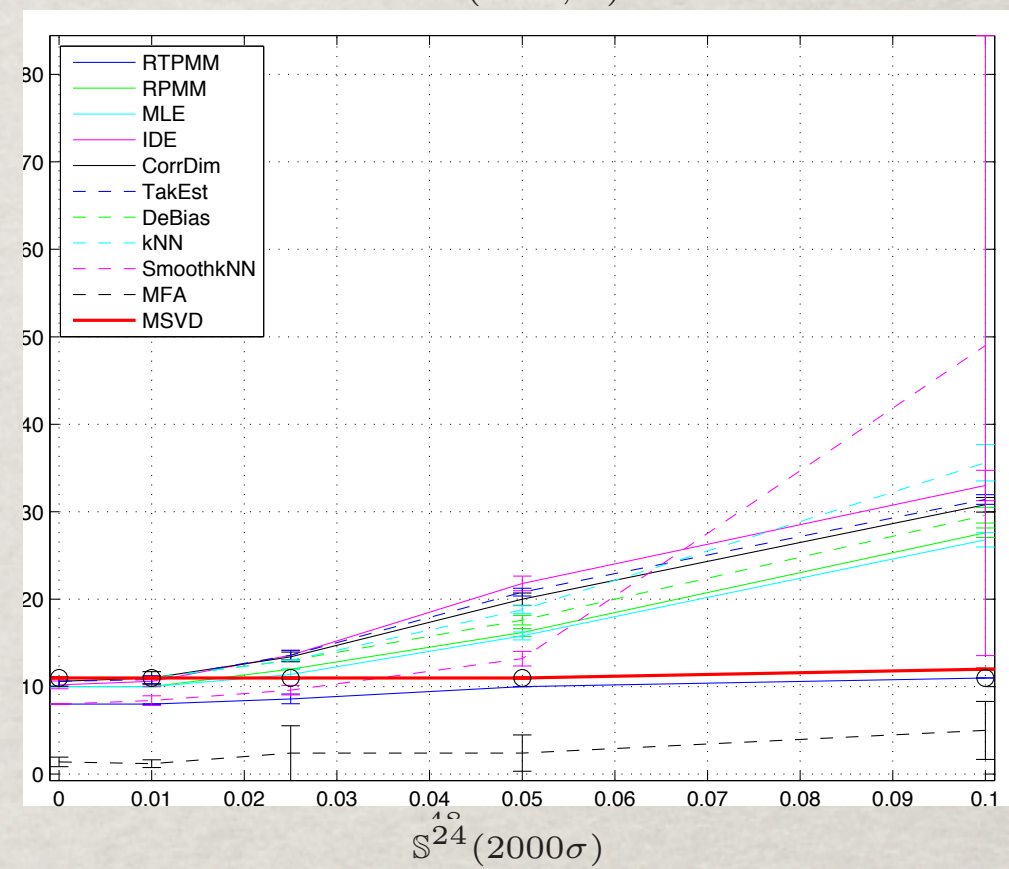


# Comparison: unit sphere

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



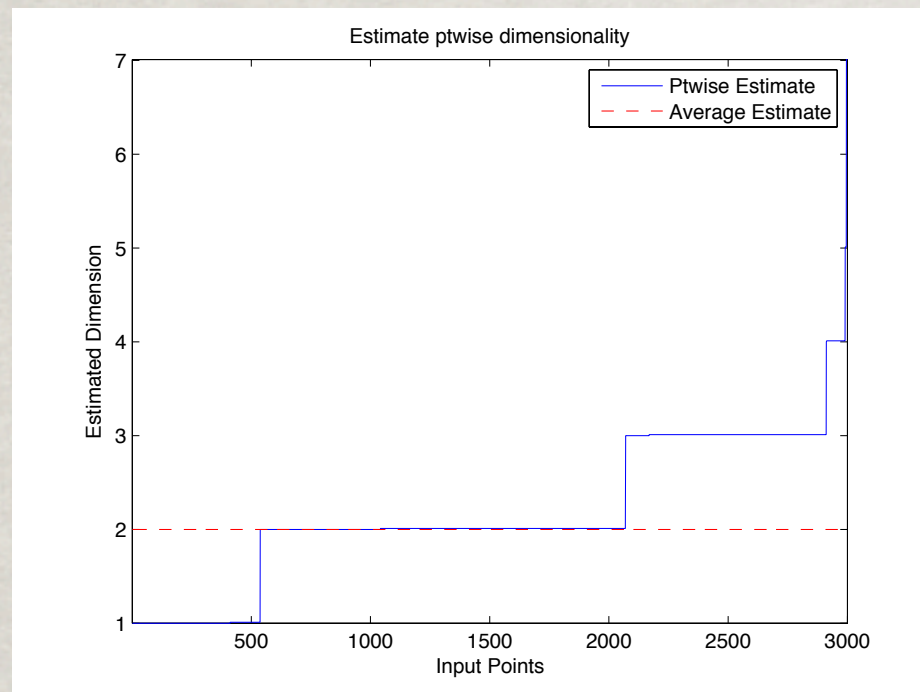
$\mathbb{S}^{12}(1000, \sigma)$



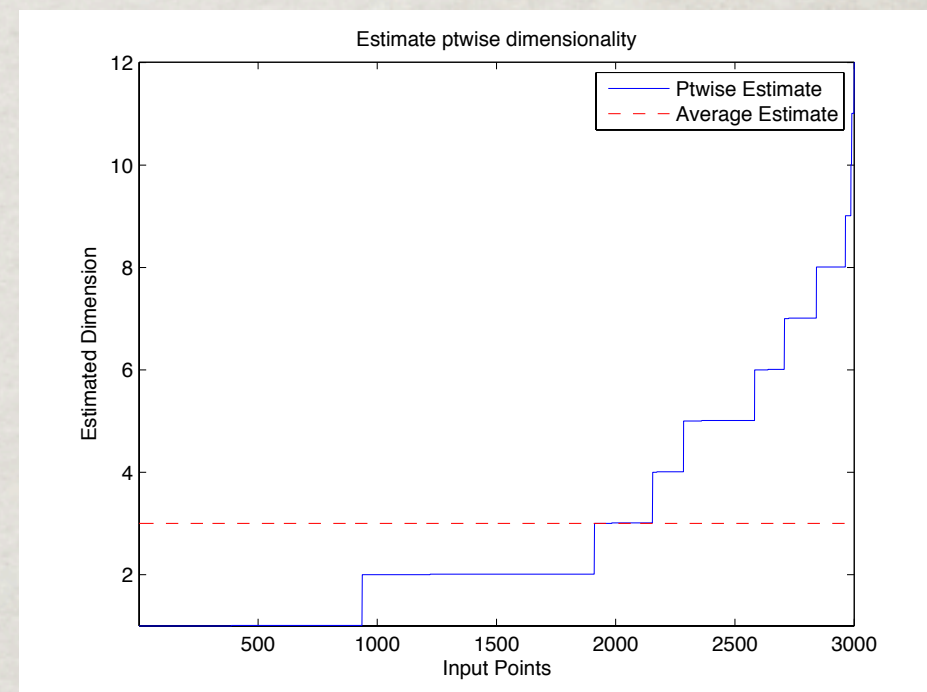


# Some data sets: heterogeneous dim.'s

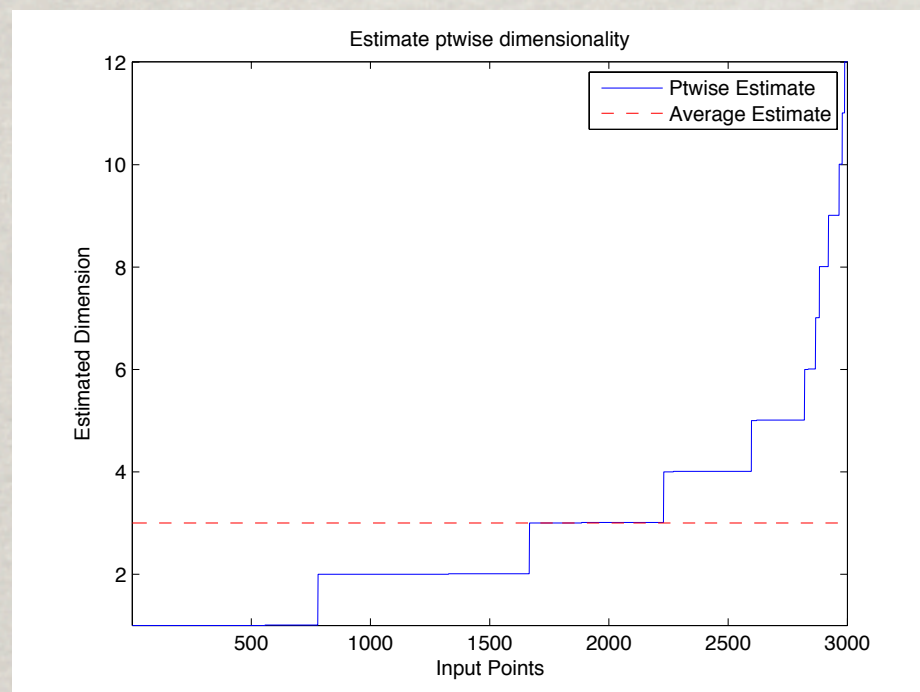
With A.V. Little, L. Rosasco, 2010



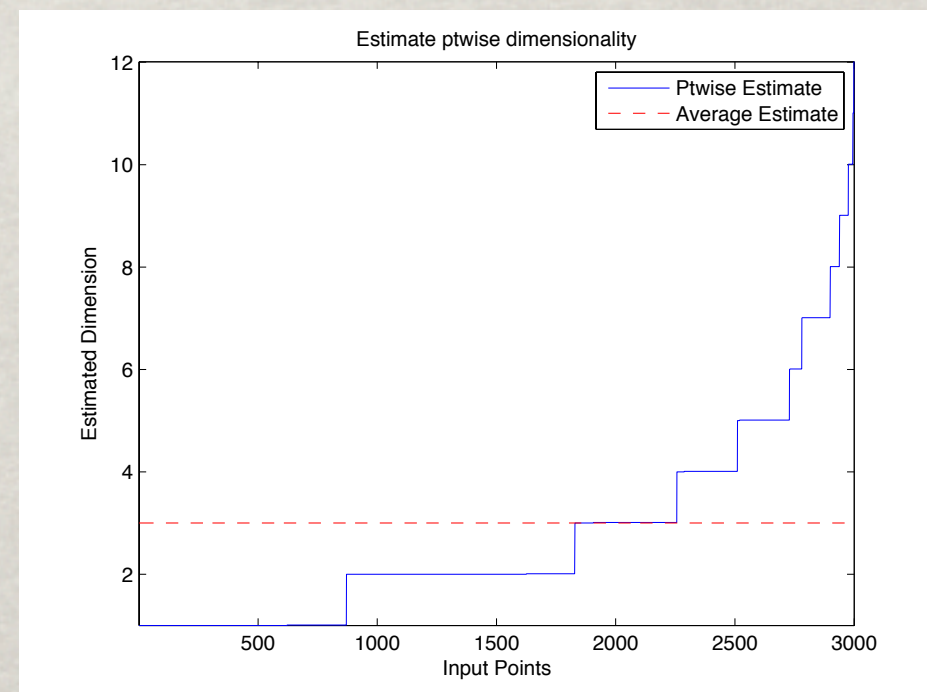
Handwritten digit 1



Handwritten digit 2



Handwritten digit 3

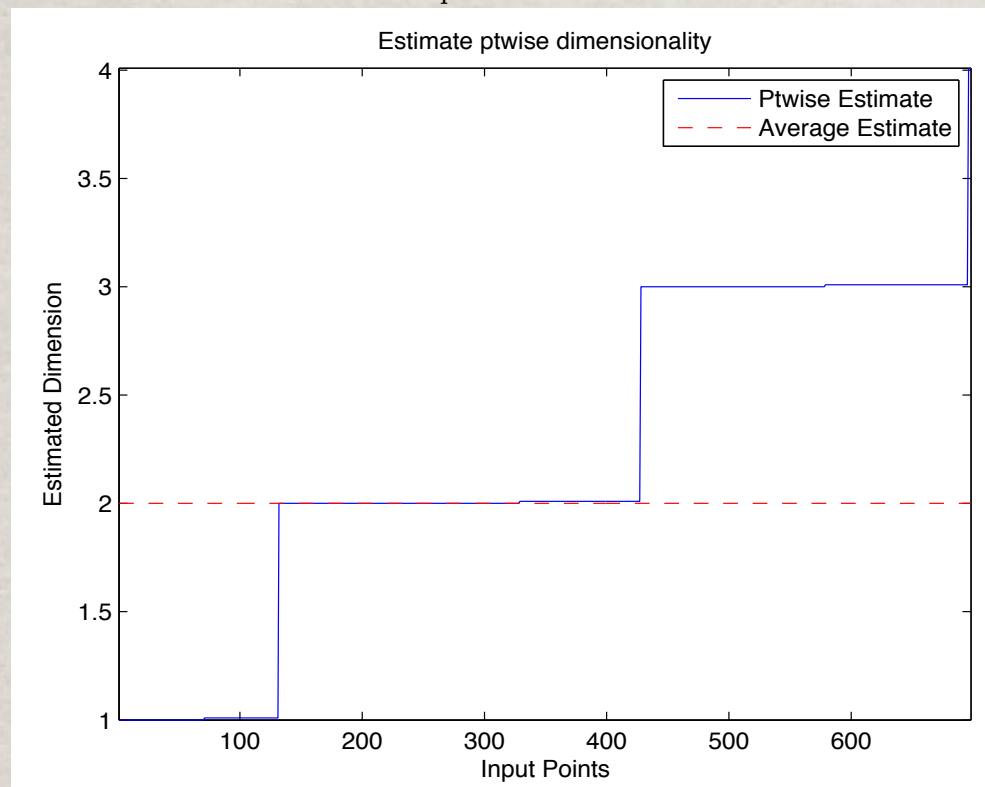


Handwritten digit 4

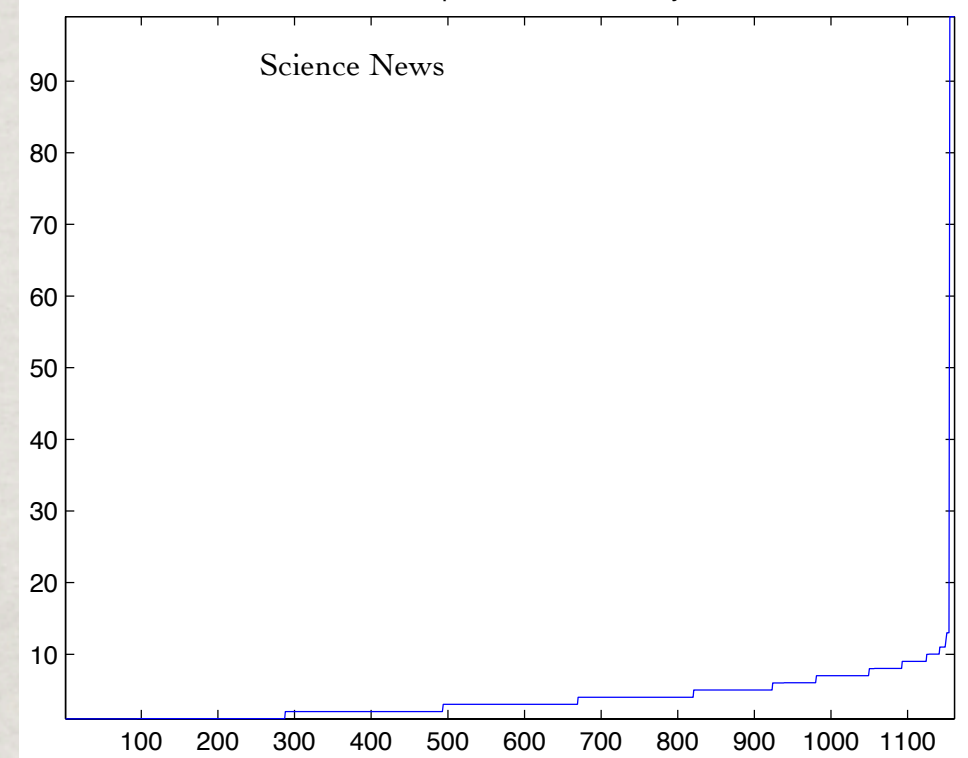


# Some data sets: heterogenous dim.'s

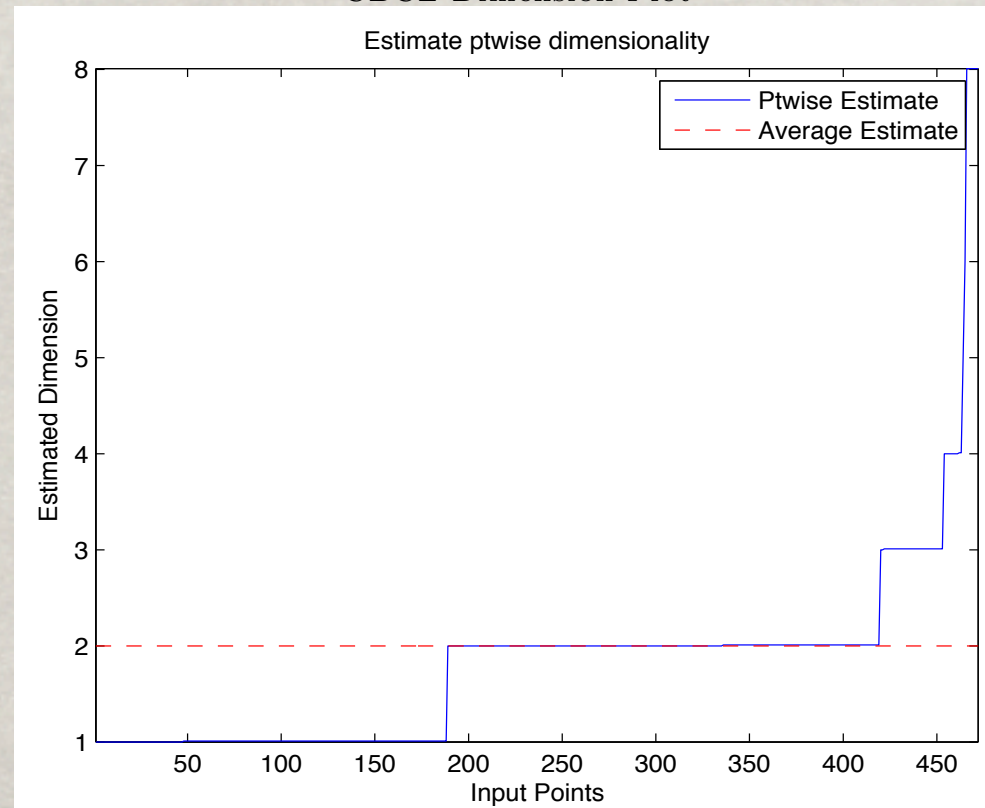
Isomap Dimension Plot



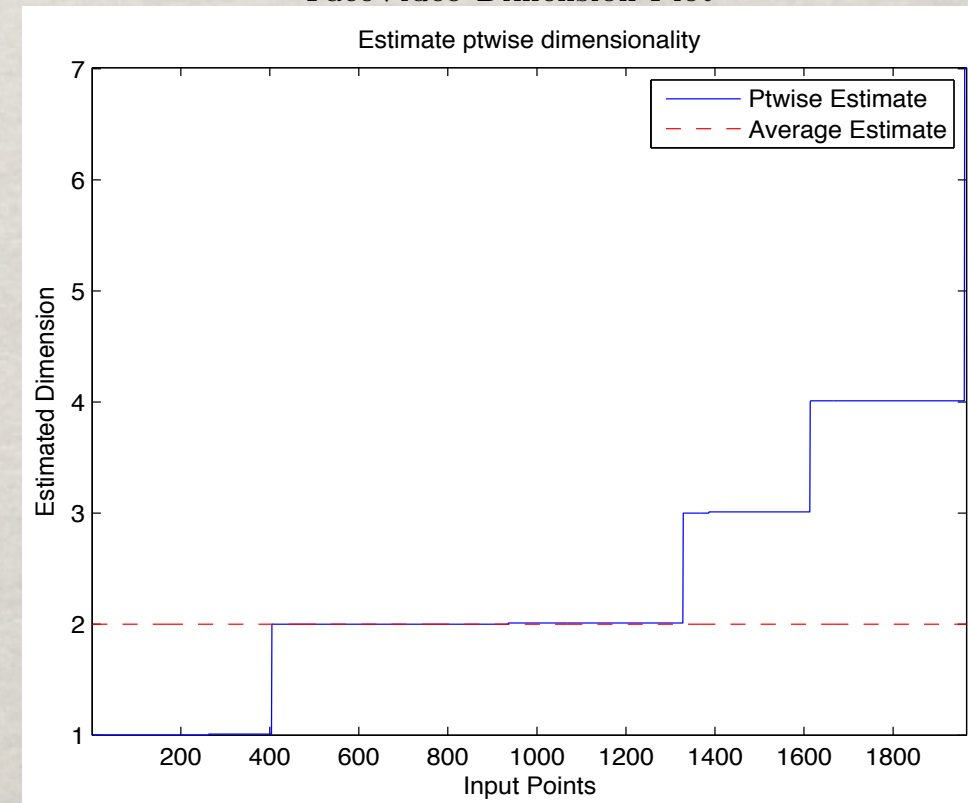
Estimate ptwise dimensionality



CBCL Dimension Plot

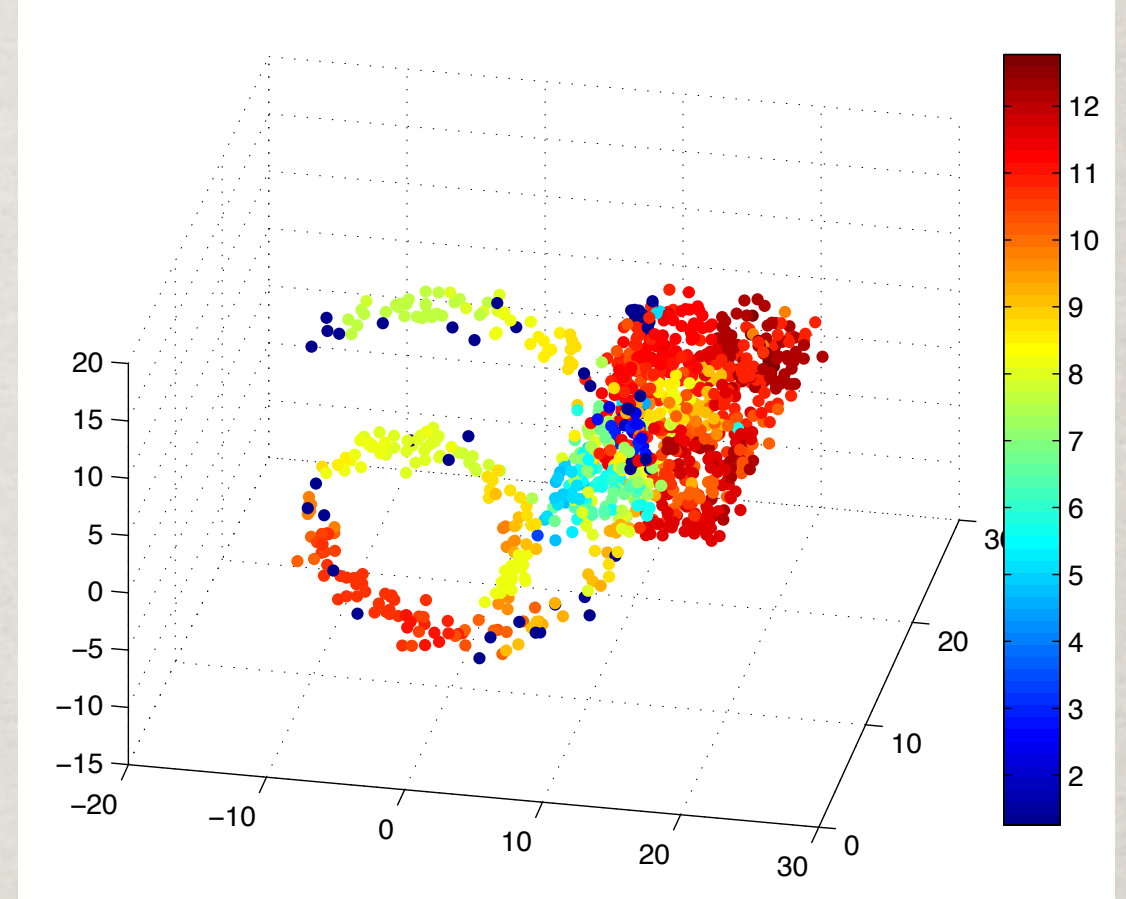
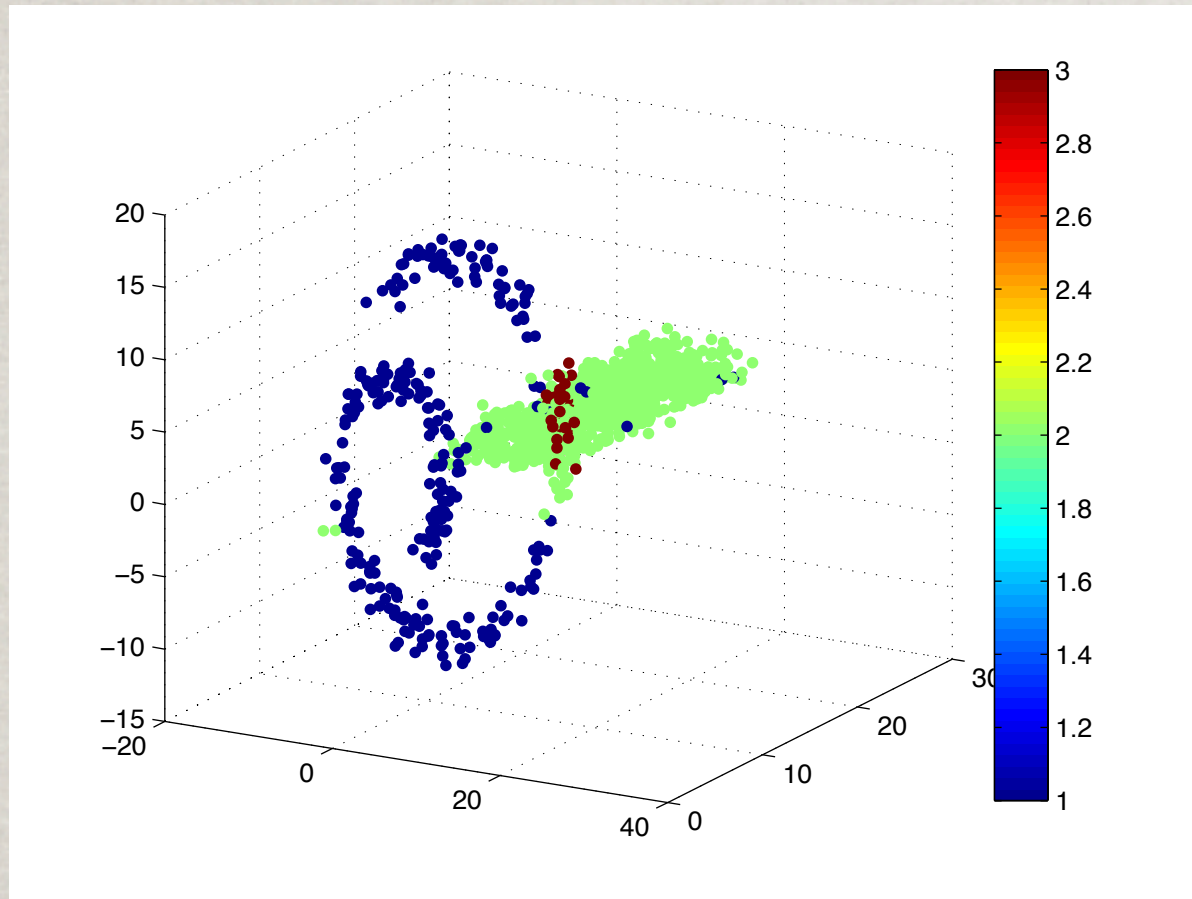


FaceVideo Dimension Plot





# Local Dim's and Scales

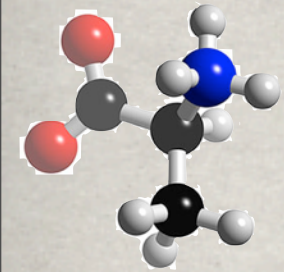


Our algorithm assigns the correct dimension dimension 1 to the spiral (because of the noise), and dimension 2 to the plane. 86% of the point of the spiral are assigned a dimension smaller than 2, and 77% of the points on the plane are assigned dimension 2 (or greater). Overall, clustering by dimension gives an accuracy of 97%, the present state-of-art to our knowledge. Bottom left: maximal good scale according to the algorithm.



# Molecular Dynamics & F-P. equation

R.R.Coifman, I.G.Kevrekidis, S.Lafon, MM, B.Nadler, *Multiscale Model. Simul.*



Fokker-Planck equation & eigenfunctions

$$\frac{\partial p}{\partial t} = - \sum_i^{3N} \frac{\partial}{\partial x_i} \left( \frac{1}{\beta} \frac{\partial}{\partial x_i} + \frac{\partial E}{\partial x_i} \right) p = -\mathbf{H}_{\text{FP}} p$$

$\beta = 1/(k_B T)$ ,  $k_B$  is Boltzmann's constant

Under suitable conditions, it has discrete spectrum  $0 = \lambda_0 < \lambda_1 \leq \dots \lambda_k \ll \lambda_{k+1} \leq \dots$ , and fundamental solution with eigen-expansion

$$p_t(x, y) = \phi_0(x) + \sum_{j=1}^{+\infty} \psi_j(y) \phi_j(x) e^{-\lambda_j t}.$$

The dual system of eigenfunctions, which we pick as reaction coordinates, is

$$\psi_j(x) = \phi_j(x) / \phi_0(x).$$

With these normalizations,

$$d^{(t)}(x, y) = ||p_t(x, \cdot) - p_t(y, \cdot)||_{L^2} = \sqrt{\sum_j e^{-\lambda_j t} |\psi_j(x) - \psi_j(y)|^2}$$



# Locally Scaled Diffusion Map

Joint with C. Clementi, M. Rohrdanz, W. Zheng, JCP 2011

- . Construct the  $N \times N$  matrix of transition probability kernels  $K$ , as

$$K_{ij} = e^{-\frac{d_{\text{RMSD}}(x_i, x_j)^2}{2\epsilon_i \epsilon_j}},$$

for  $x_i$  and  $x_j$  molecular configurations,  $\epsilon_i$  and  $\epsilon_j$  their local scales.

- . For each  $x_i$ , compute

$$P_i = \sum_{j=1}^N K_{ij},$$

which is proportional to a density estimation around  $x_i$ .

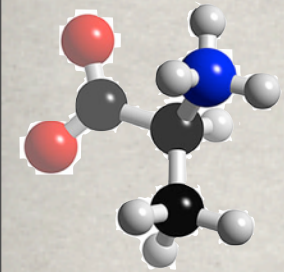
- . Normalize the kernel as

$$\tilde{K}_{ij} = P_i^{-\frac{1}{2}} K_{ij} P_j^{-\frac{1}{2}}.$$

- . Define the diagonal matrix  $D$  as  $D_i = \sum_{j=1}^N \tilde{K}_{ij}$ , and construct a Markov matrix  $M = D^{-1} \tilde{K}$ ,

$$M_{ij} = D_i^{-1} \tilde{K}_{ij}.$$

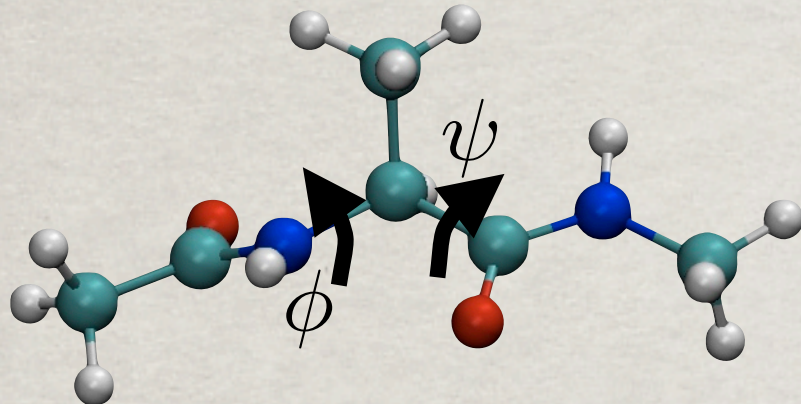
- . Compute largest eigenvalues and corresponding right eigenvectors of  $M$ .



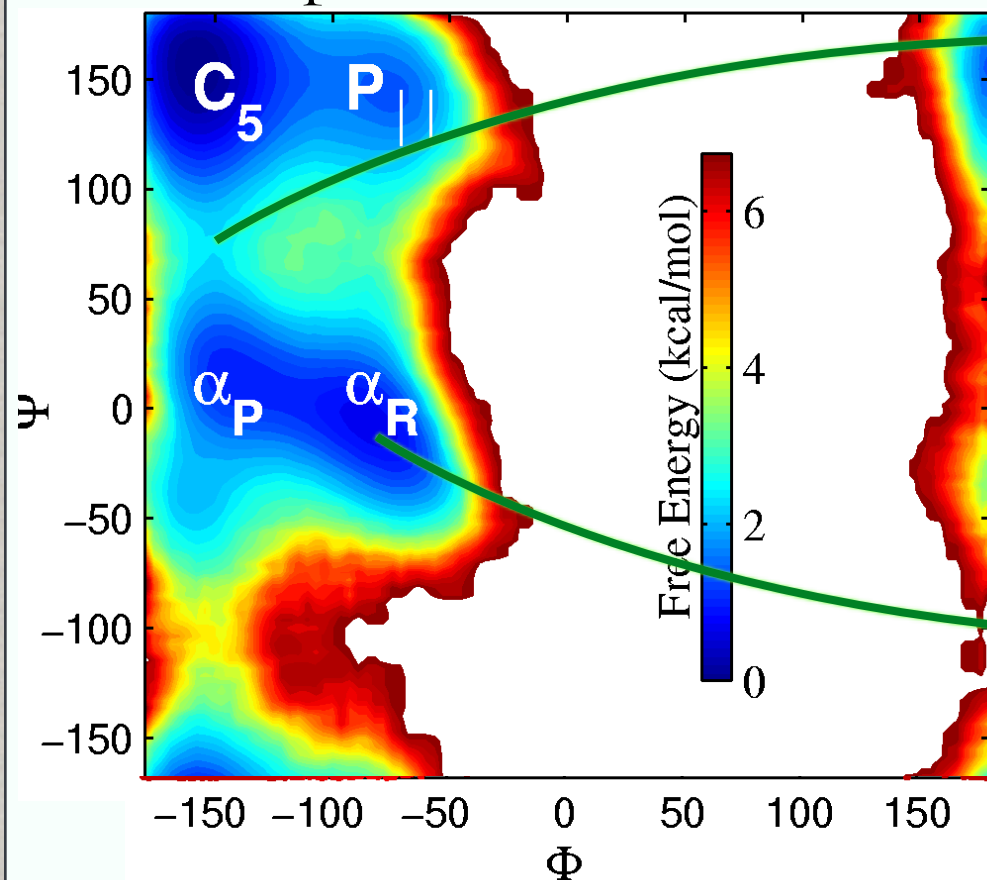


# Example: Alanine dipeptide

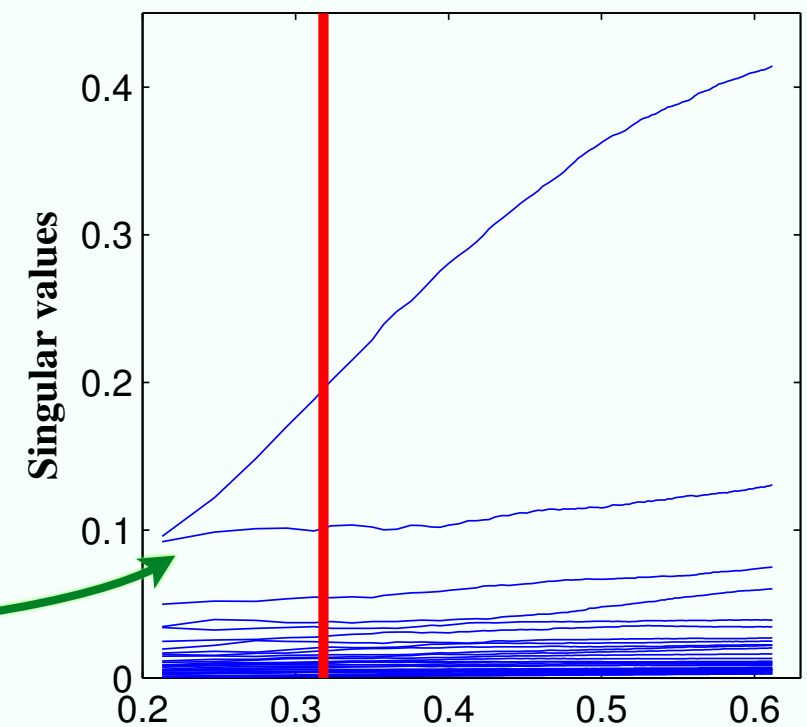
Joint with C. Clementi, M. Rohrdanz, W. Zheng, JCP 2011



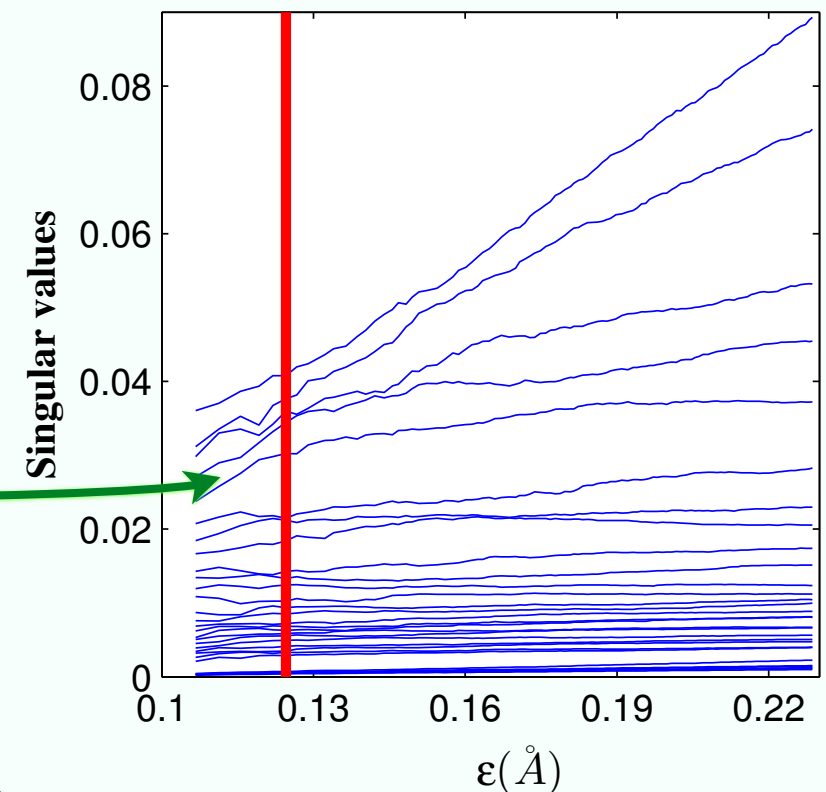
Free energy in terms of  
empirical coordinates



MSVD near  
transition state



MSVD near free  
energy minimum

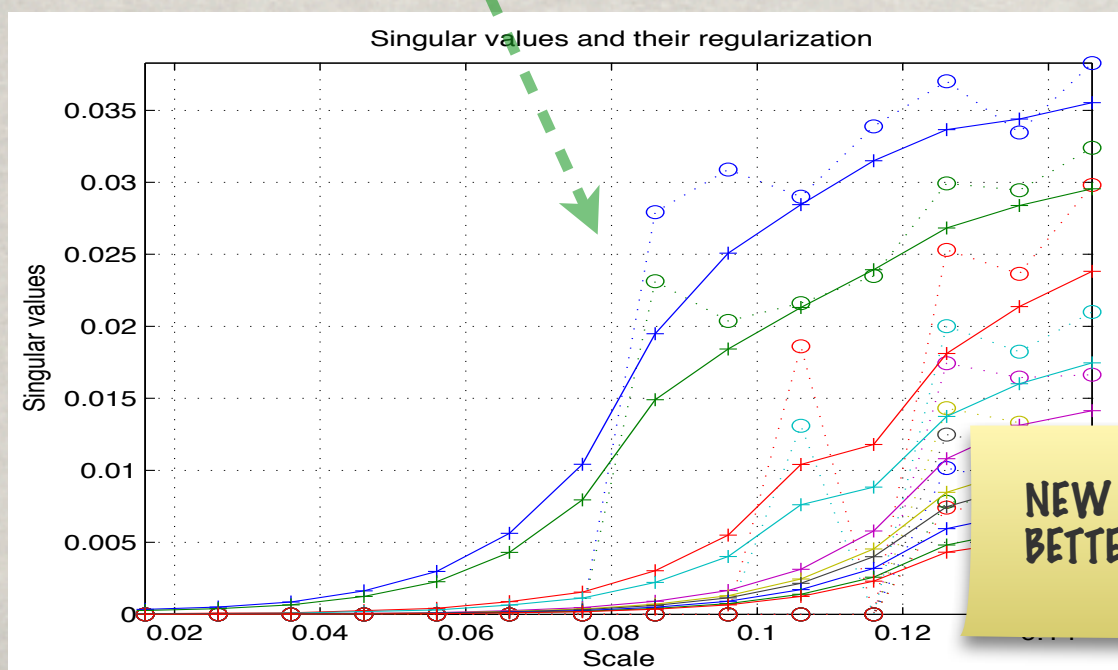
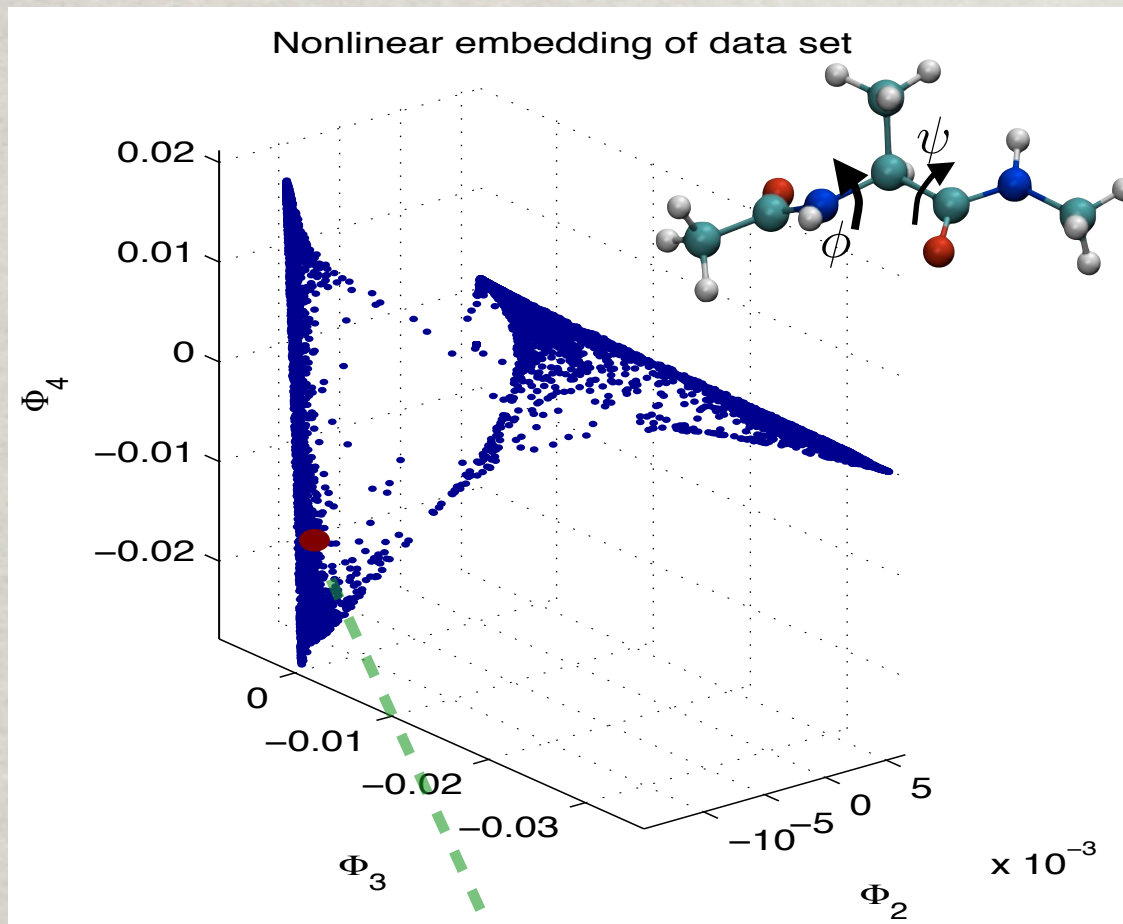




# Molecular Dynamics data for alanine

Joint with C. Clementi, M. Rohrdanz, W. Zheng, JCP 2011

Data points are configurations of alanine dipeptide in water bath at constant temperature.



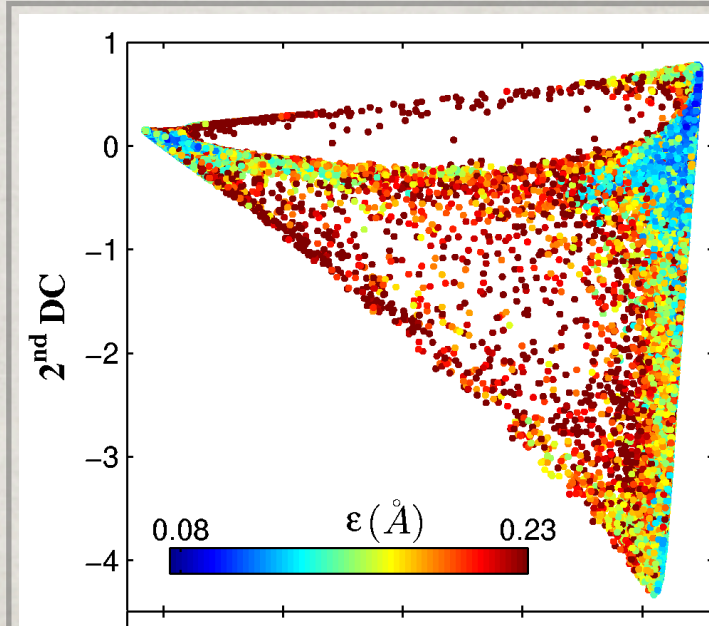
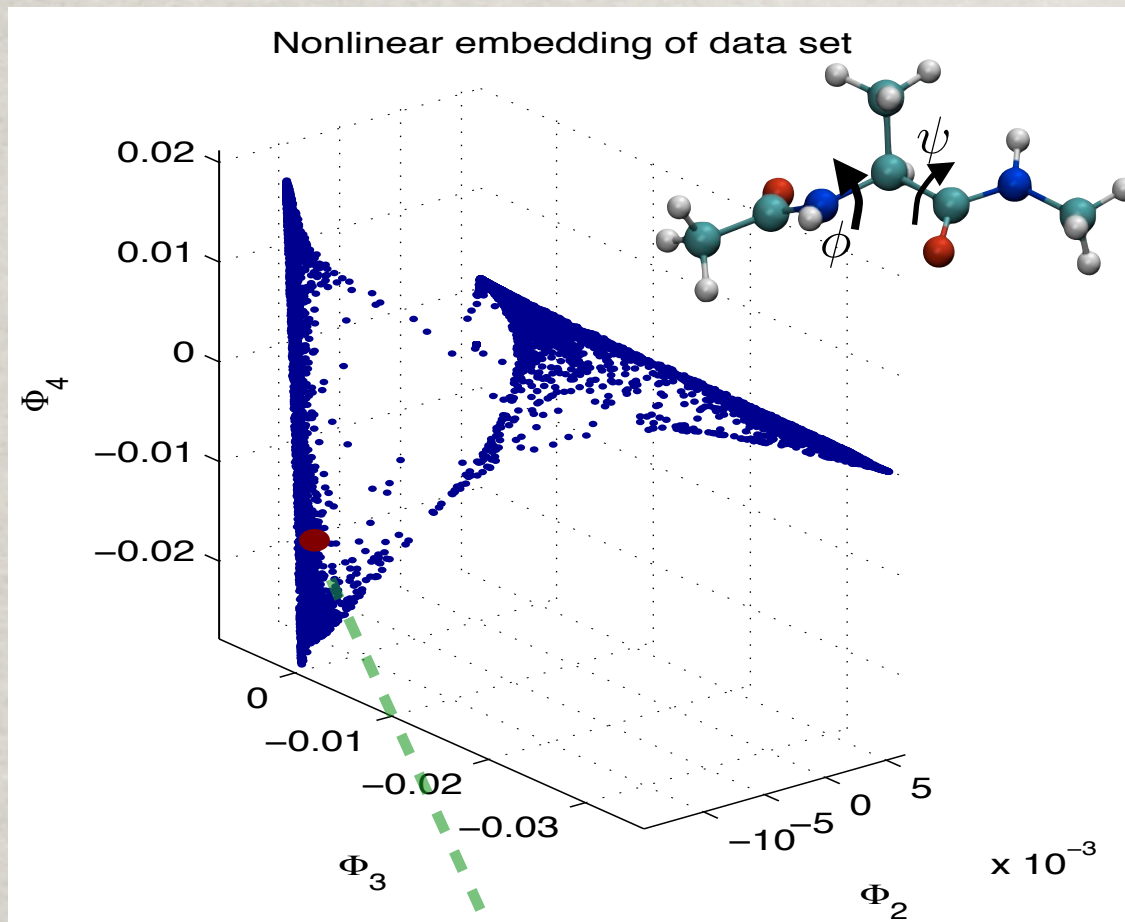
NEW type of diffusion maps THANKS to this  
BETTER STATISTICS for the DYNAMICS



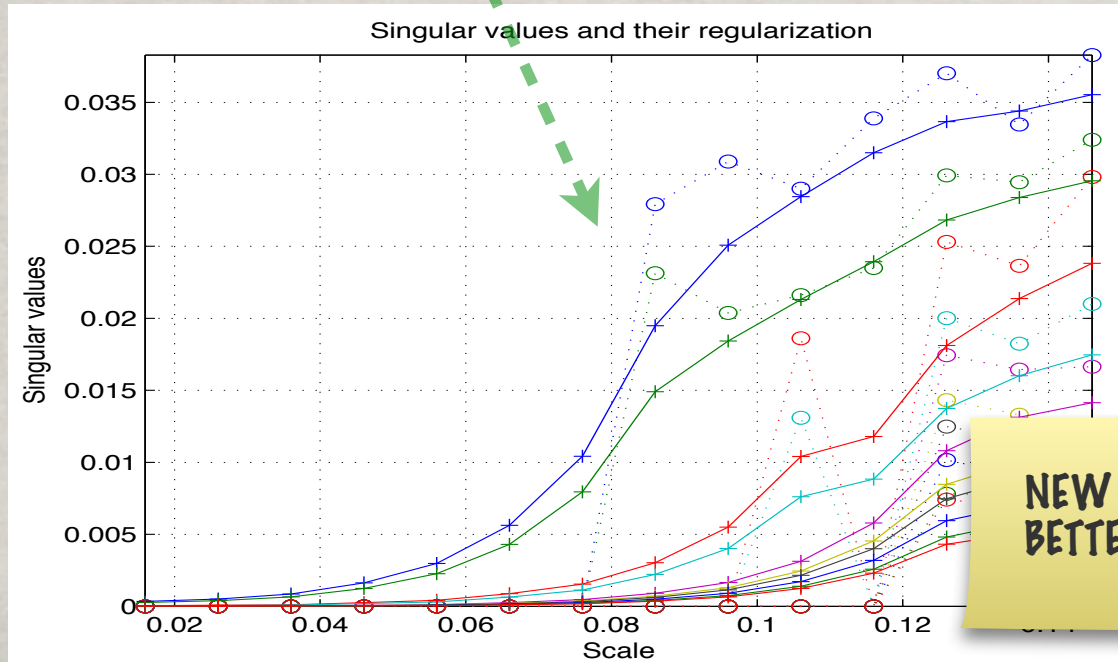
# Molecular Dynamics data for alanine

Joint with C. Clementi, M. Rohrdanz, W. Zheng, JCP 2011

Data points are configurations of alanine dipeptide in water bath at constant temperature.



We use multiscale singular values to detect a natural local scale of the data. Different regions of the effective state space do exhibit very different local scales.



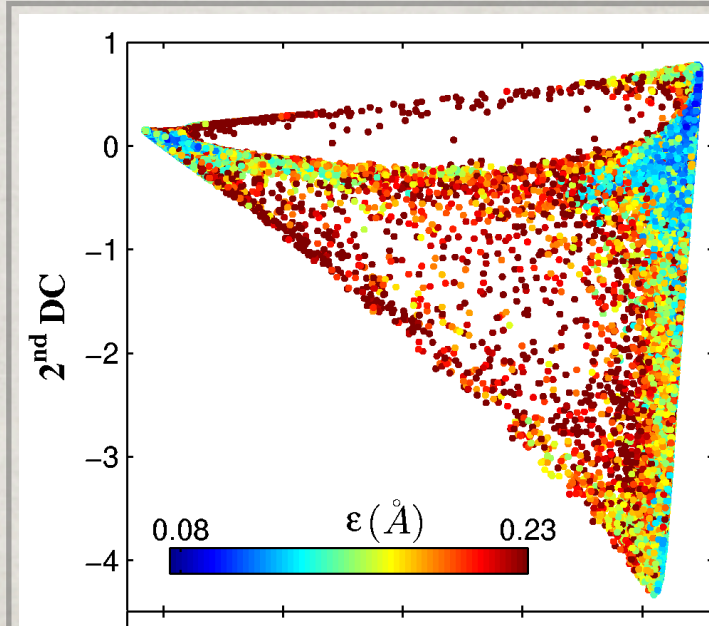
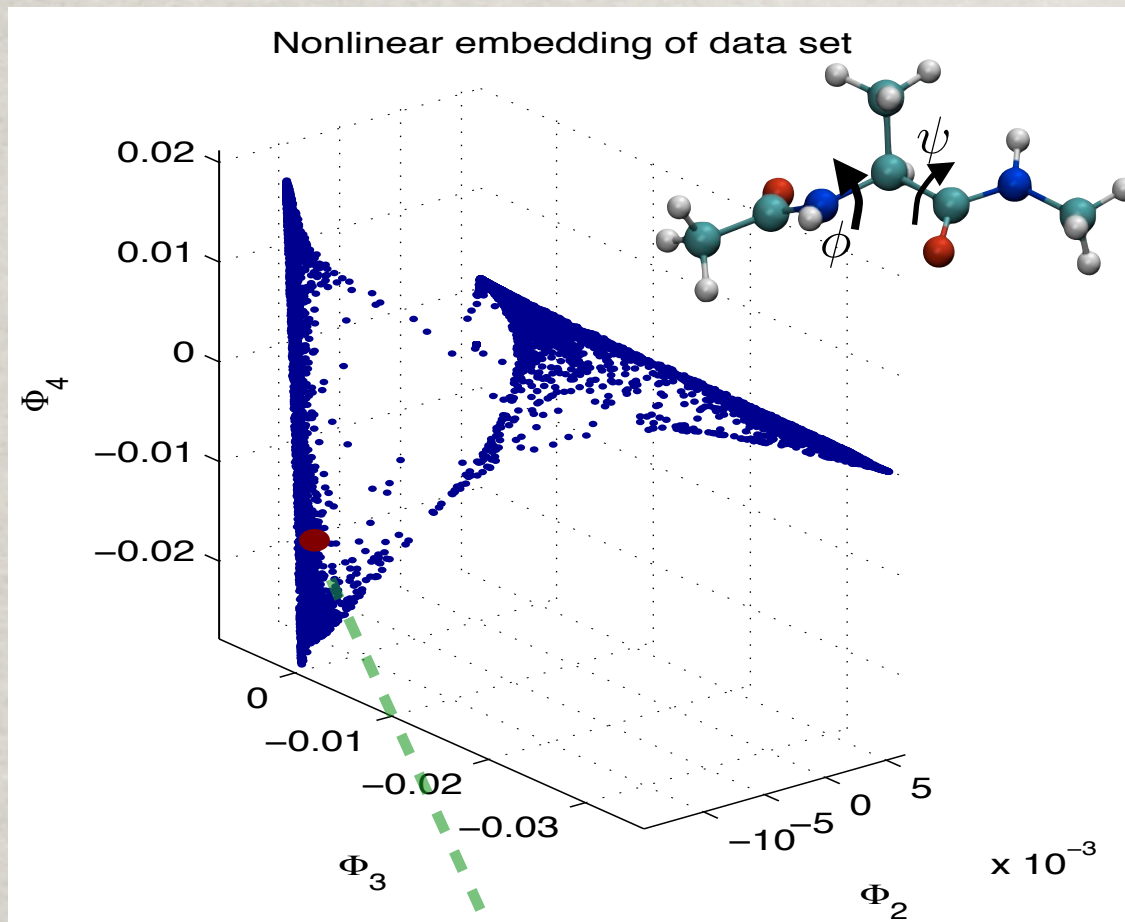
NEW type of diffusion maps THANKS to this  
BETTER STATISTICS for the DYNAMICS



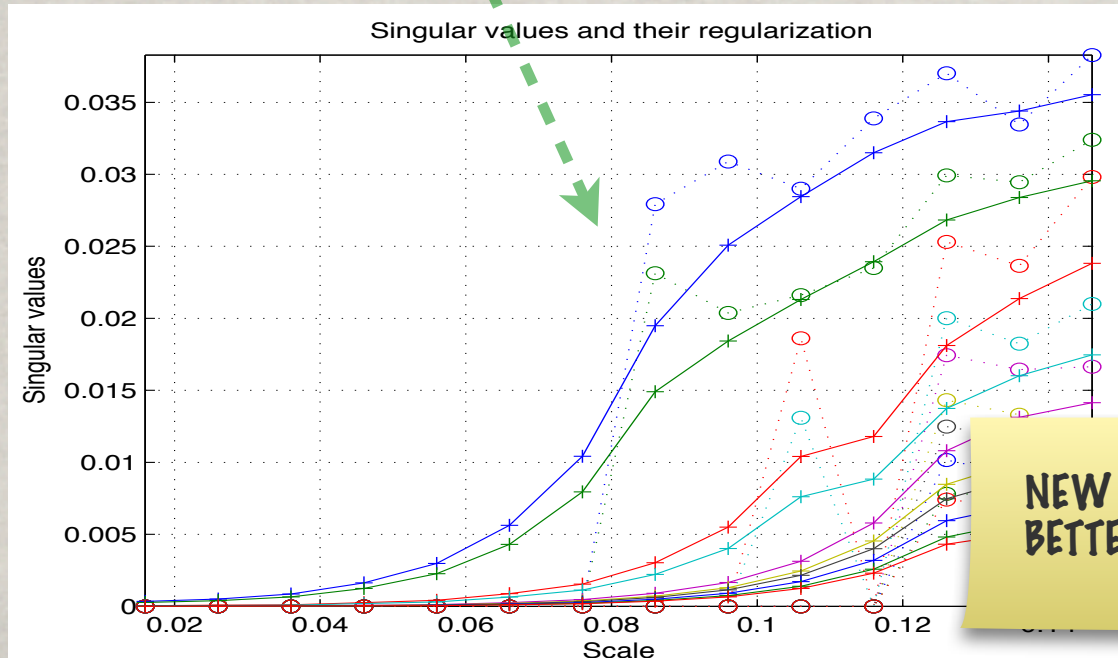
# Molecular Dynamics data for alanine

Joint with C. Clementi, M. Rohrdanz, W. Zheng, JCP 2011

Data points are configurations of alanine dipeptide in water bath at constant temperature.

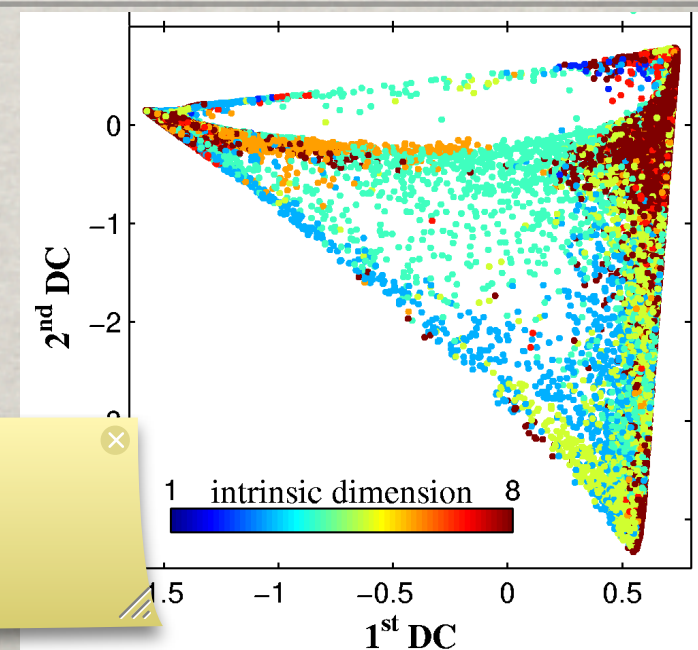


We use multiscale singular values to detect a natural local scale of the data. Different regions of the effective state space do exhibit very different local scales.



Different regions of the state space have different intrinsic dimension.

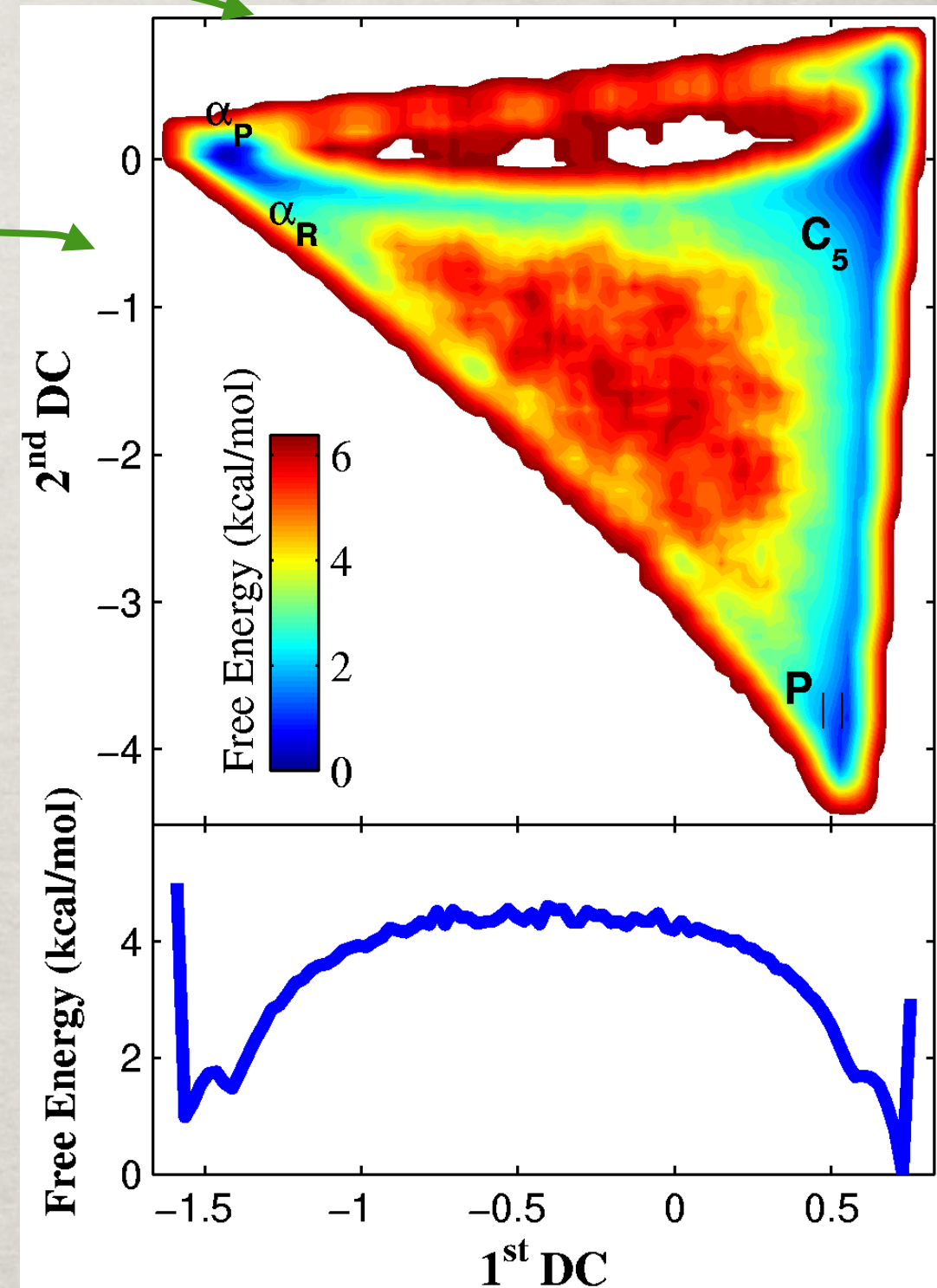
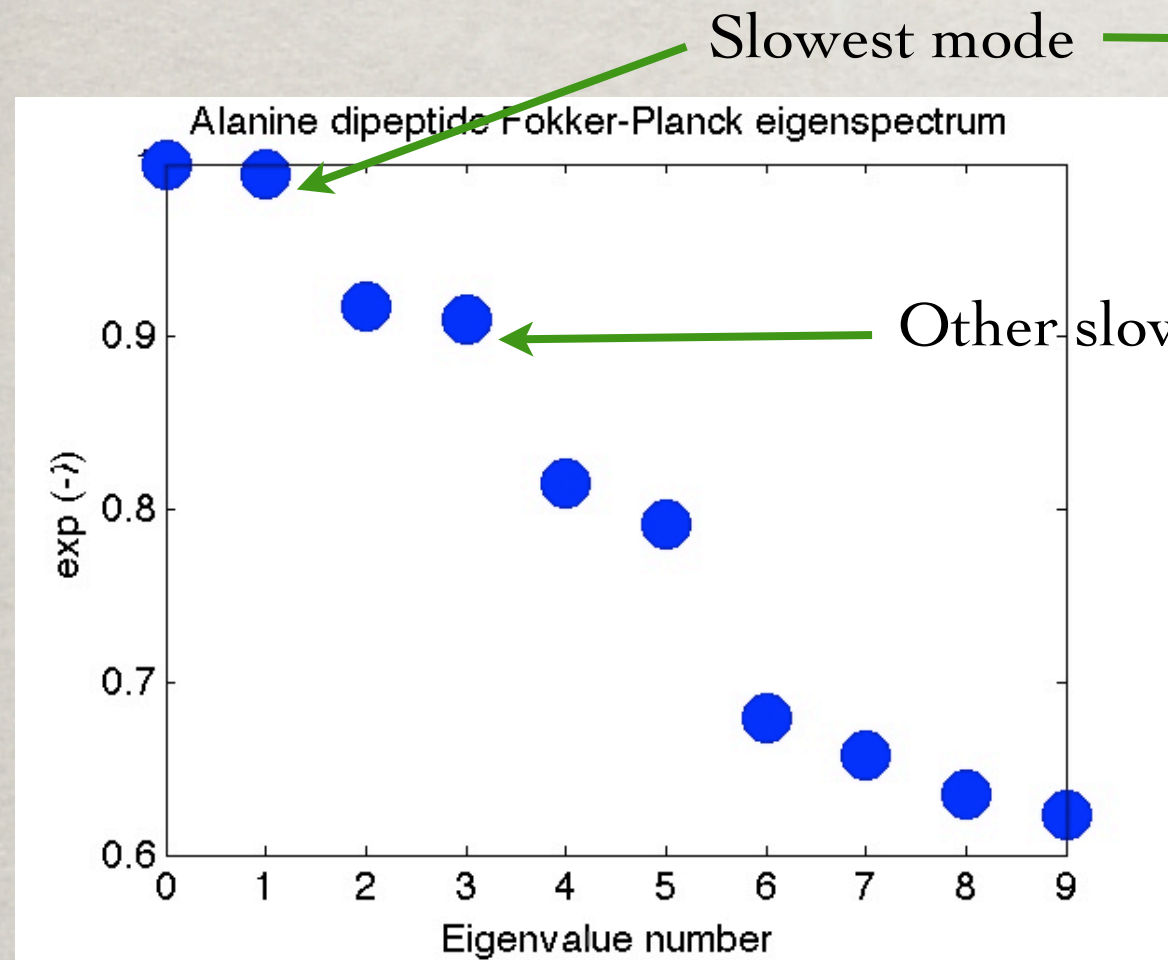
**NEW** type of diffusion maps THANKS to this  
**BETTER STATISTICS** for the DYNAMICS





# Molecular Dynamics data for alanine

Joint with C. Clementi, M. Rohrdanz, W. Zheng, JCP 2011



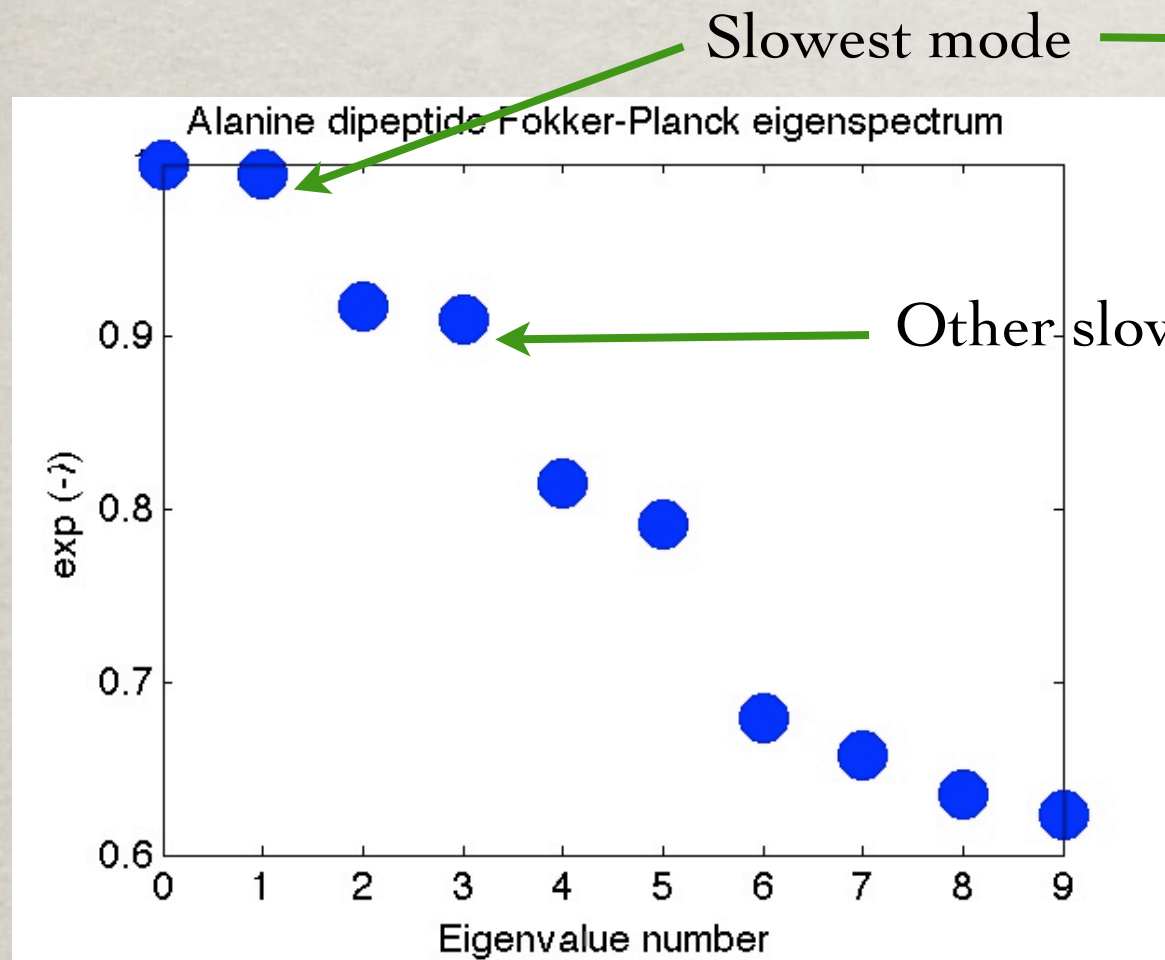
$$\text{rate} = \left( \int_{\text{barrier}} \frac{e^{\beta F(x)}}{D(x)} dx \int_{\text{well}} e^{-\beta F(x')} dx' \right)^{-1}$$

G. Hummer, 2005



# Molecular Dynamics data for alanine

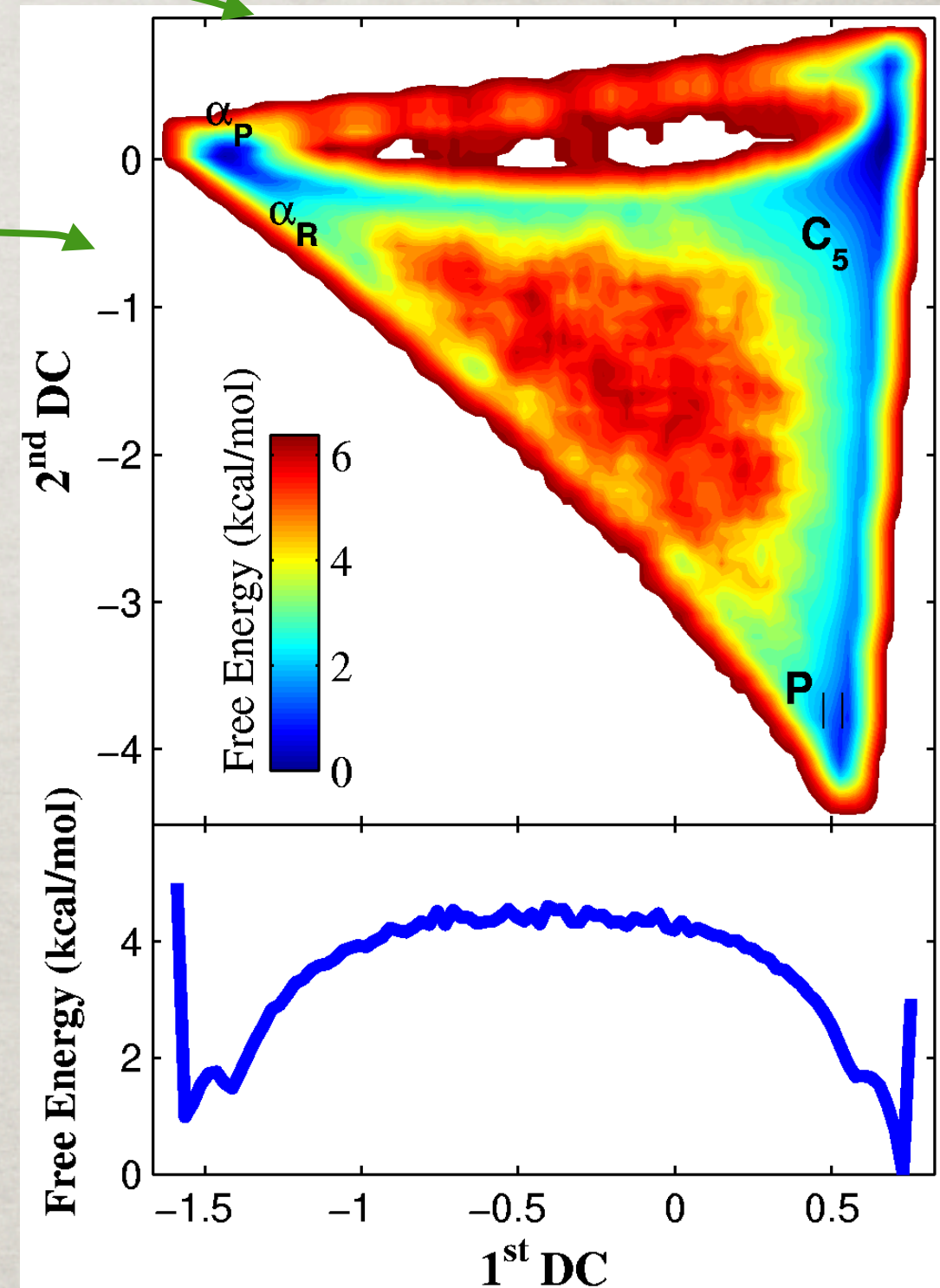
Joint with C. Clementi, M. Rohrdanz, W. Zheng, JCP 2011



Coordinate	$C_5, P_{  } \rightarrow \alpha_R \alpha_P$	$\alpha_R \alpha_P \rightarrow C_5, P_{  }$
From simulation <sup>b</sup>	0.023	0.047
1 <sup>st</sup> DC	$0.023 \pm 0.001$	$0.048 \pm 0.003$
$\Psi$	$0.020 \pm 0.001$	$0.040 \pm 0.003$

$$\text{rate} = \left( \int_{\text{barrier}} \frac{e^{\beta F(x)}}{D(x)} dx \int_{\text{well}} e^{-\beta F(x')} dx' \right)^{-1}$$

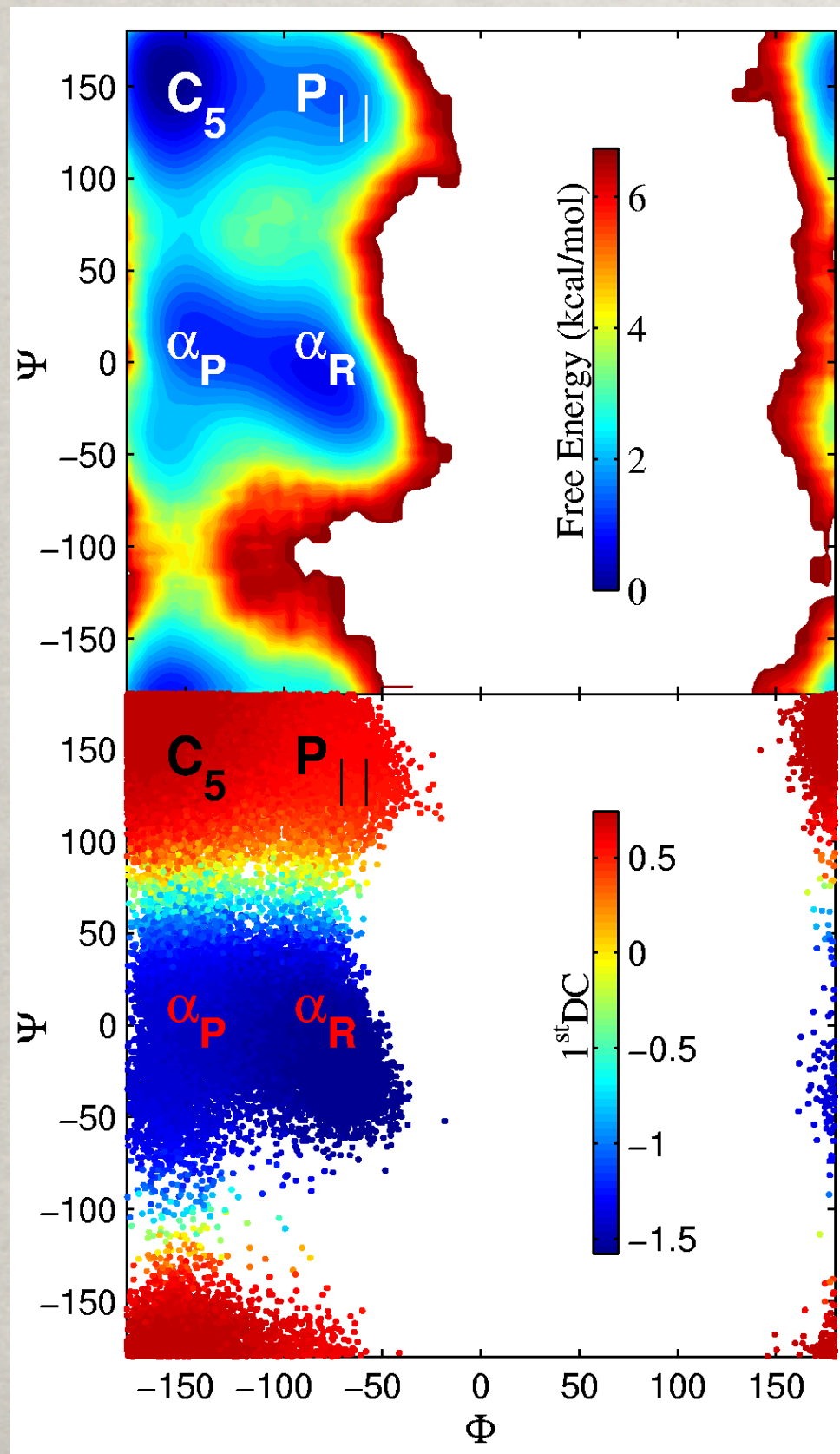
G. Hummer, 2005



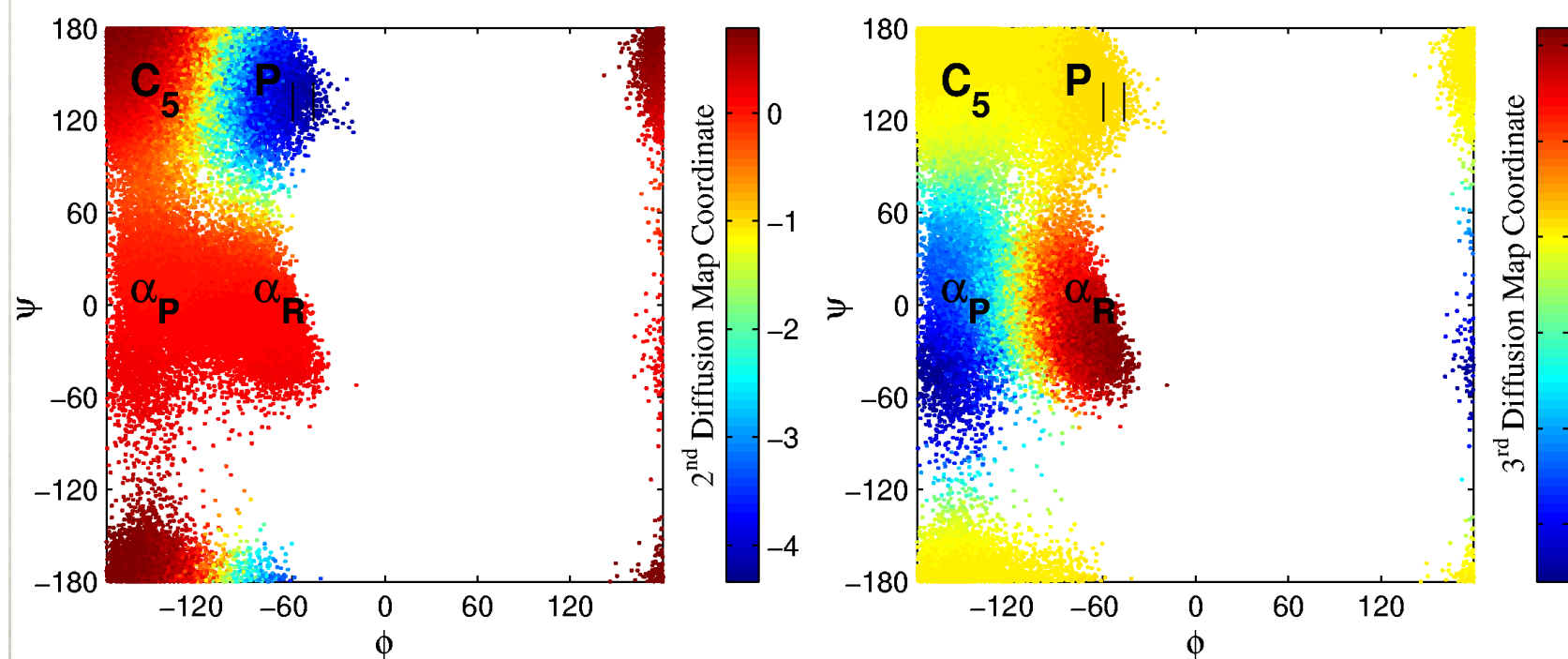


# Diffusion coord.'s - empirical coord.'s

Joint with C. Clementi, M. Rohrdanz, W. Zheng, JCP 2011



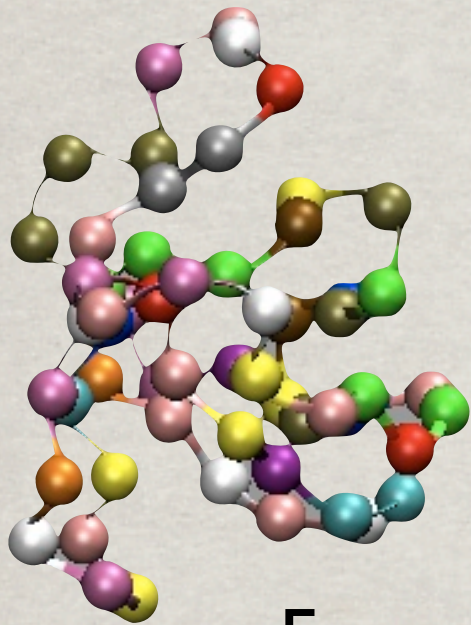
We may plot the diffusion coordinates as functions of the physical observables given by the angles and notice they are essentially in one-to-one correspondence, with the diffusion coordinates emphasizing energy barriers separating minima.



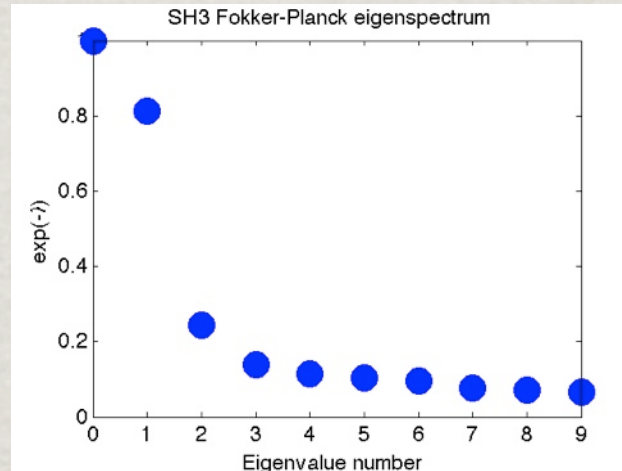


# Example: SH-3

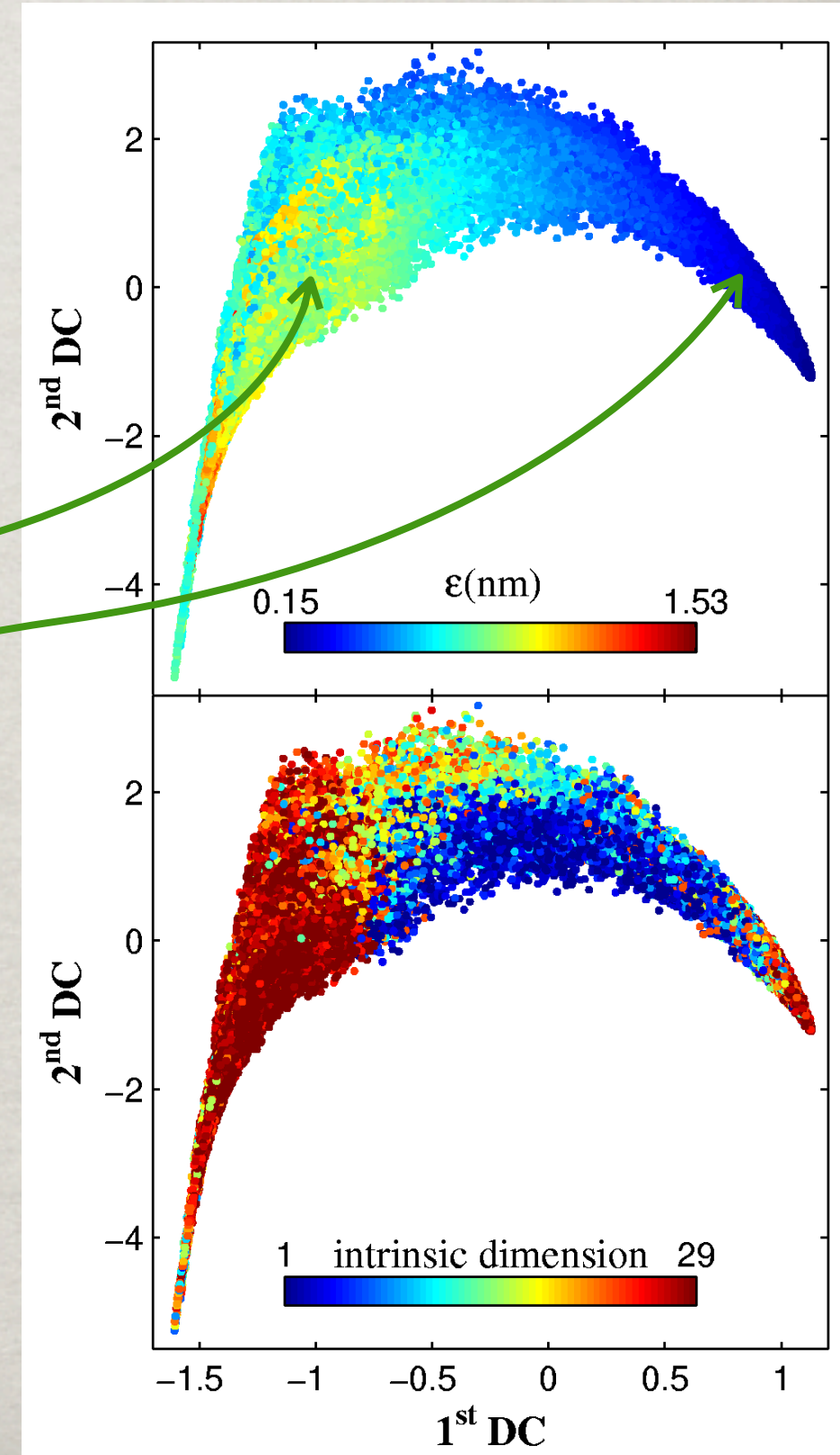
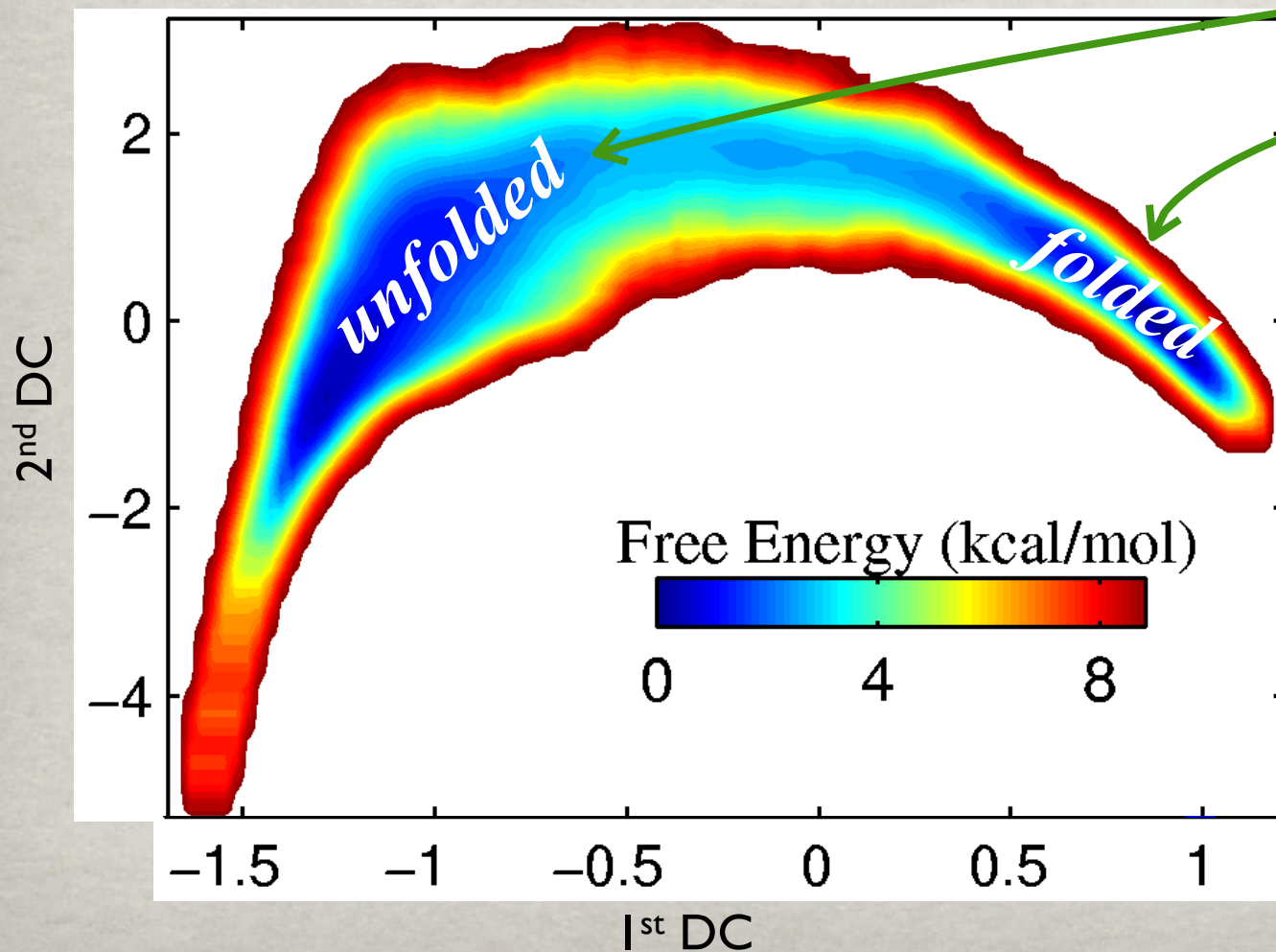
Joint with C. Clementi, M. Rohrdanz, W. Zheng, JCP 2011



$$e^{-\lambda_i}$$



Free energy in terms of  
diffusion coordinates





# Example: b3s

C. Clementi, M. Rohrdanz, W. Zheng, ongoing

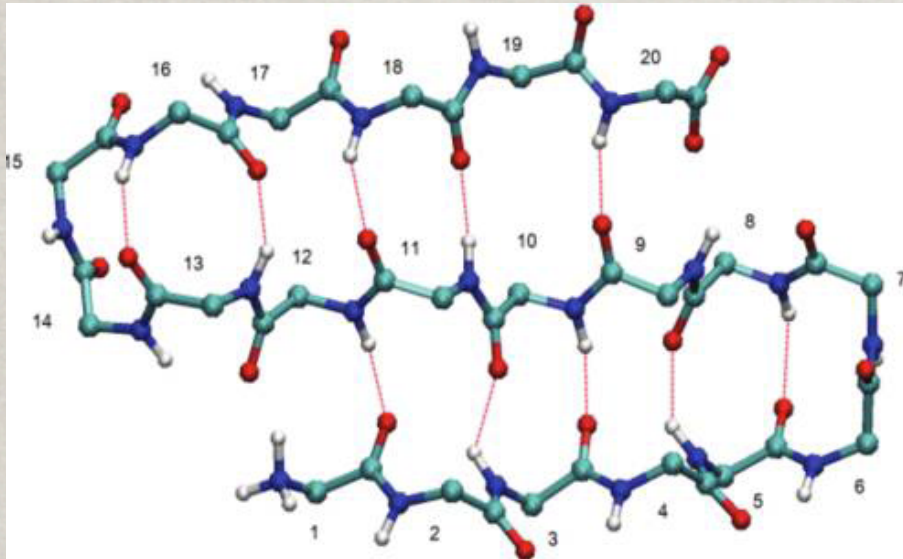
20-residue antiparallel  $\beta$ -sheet miniprotein (Beta3s)

169 non-hydrogen atoms

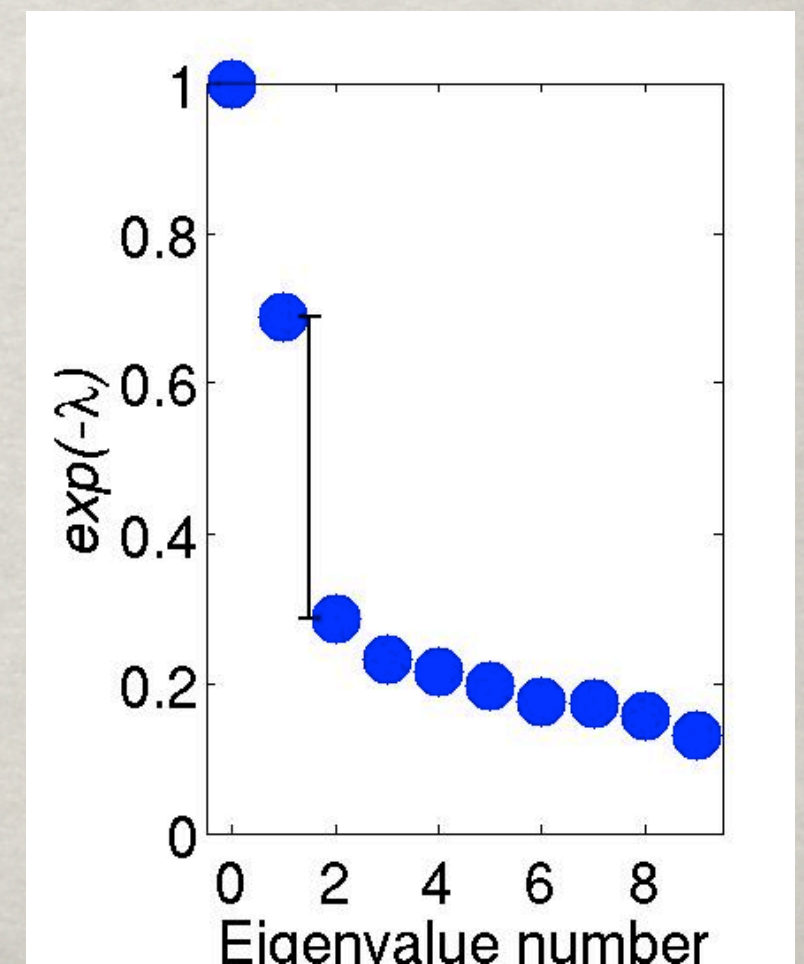
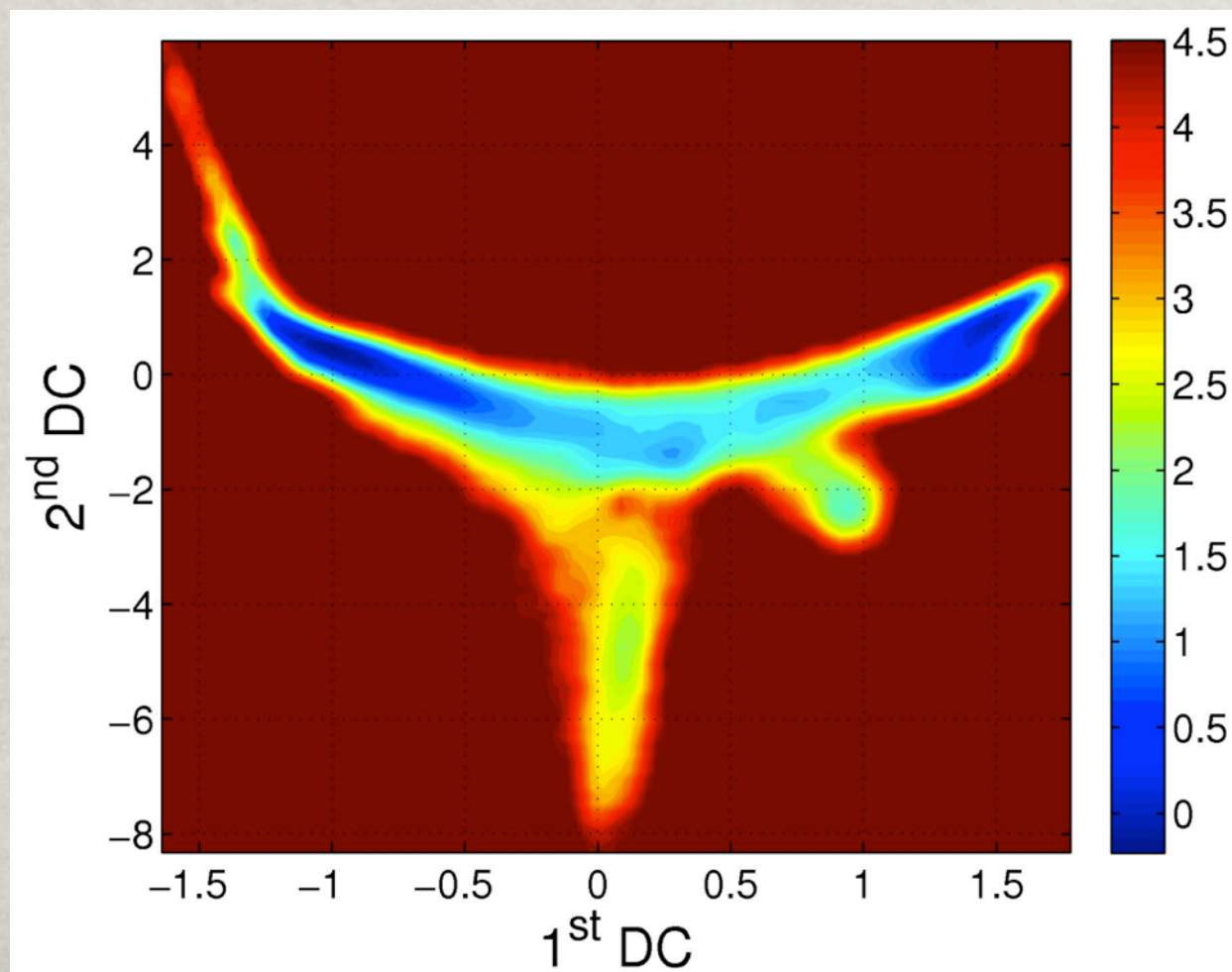
507 degrees of freedom

Previous work from other groups:

- Clustering method
- Genetic neural network approach



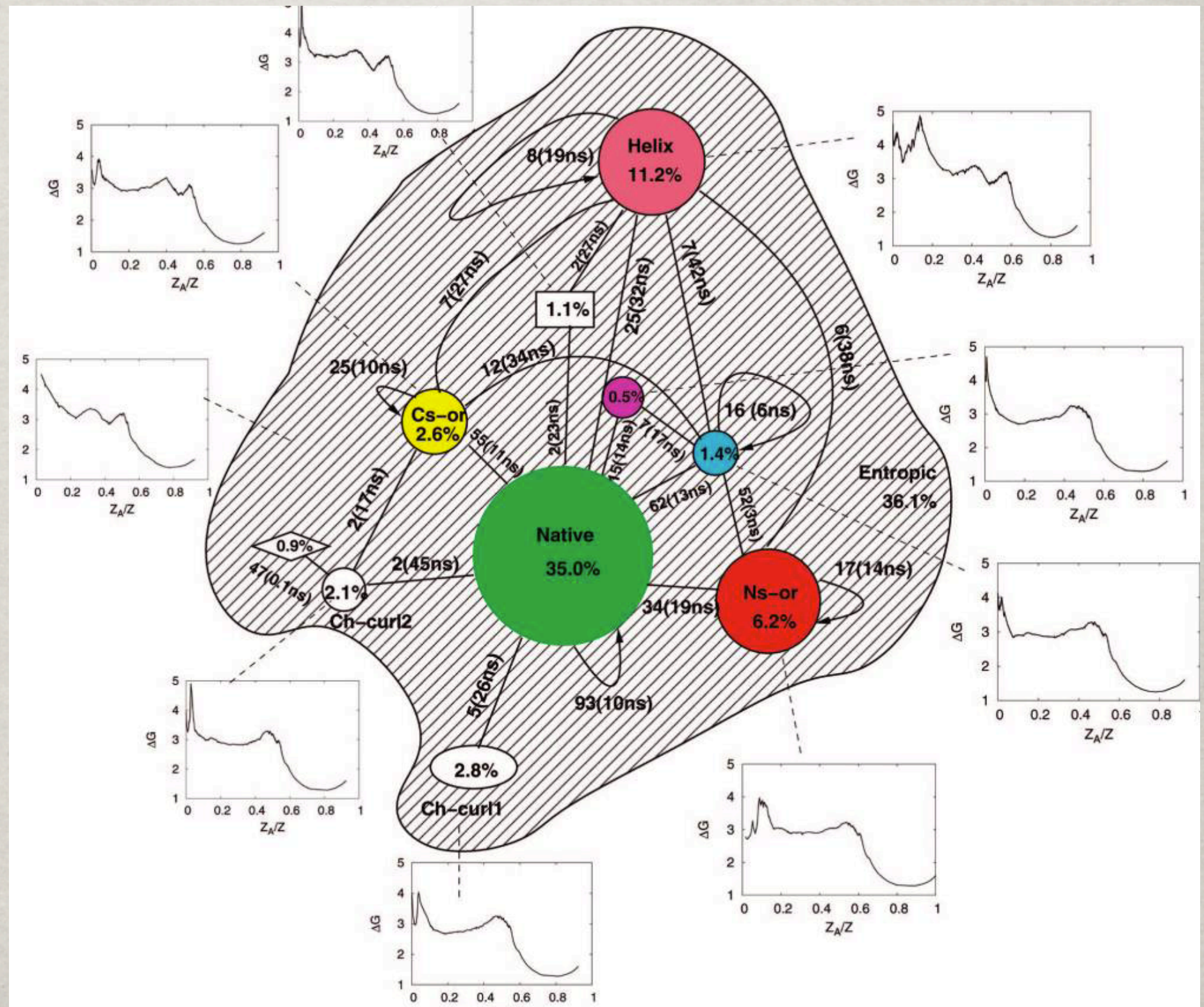
Qi B., Muff S., Caflisch A., Dinner A.,  
J. Phys. Chem. B 2010, 114, 6979





# Example: b3s

Clustering by using  
secondary structure  
sequence

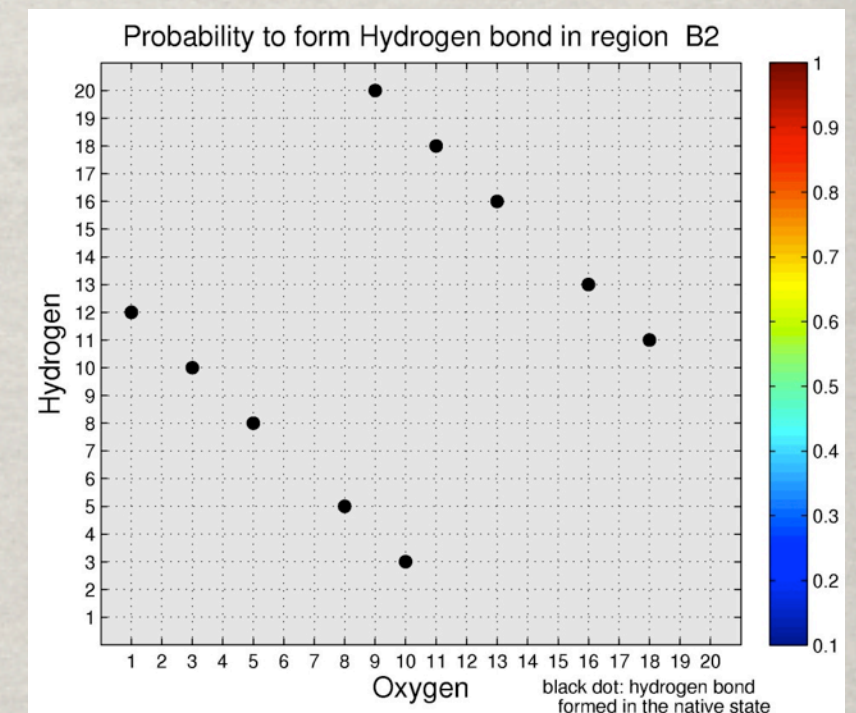
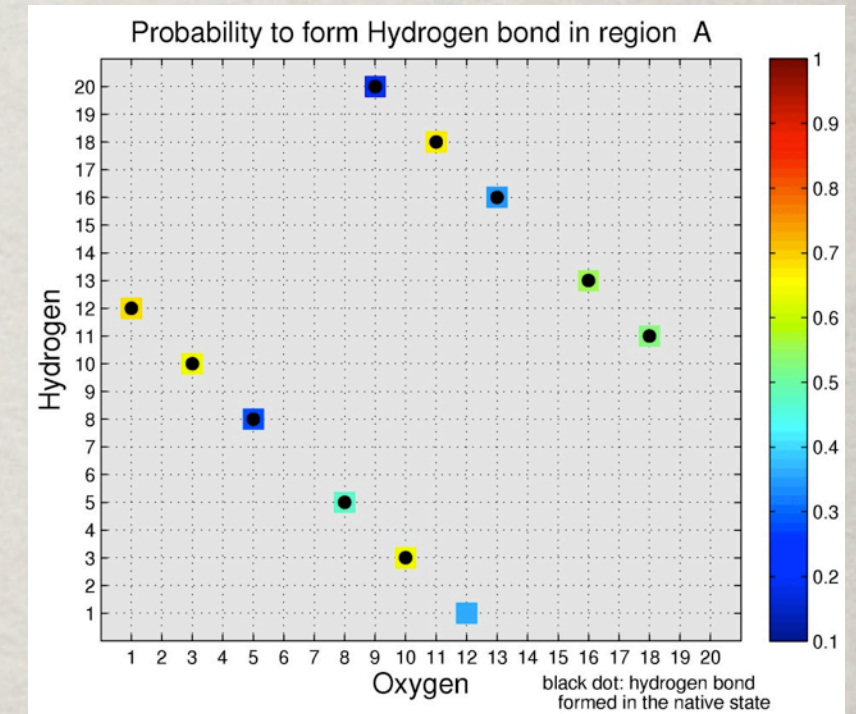
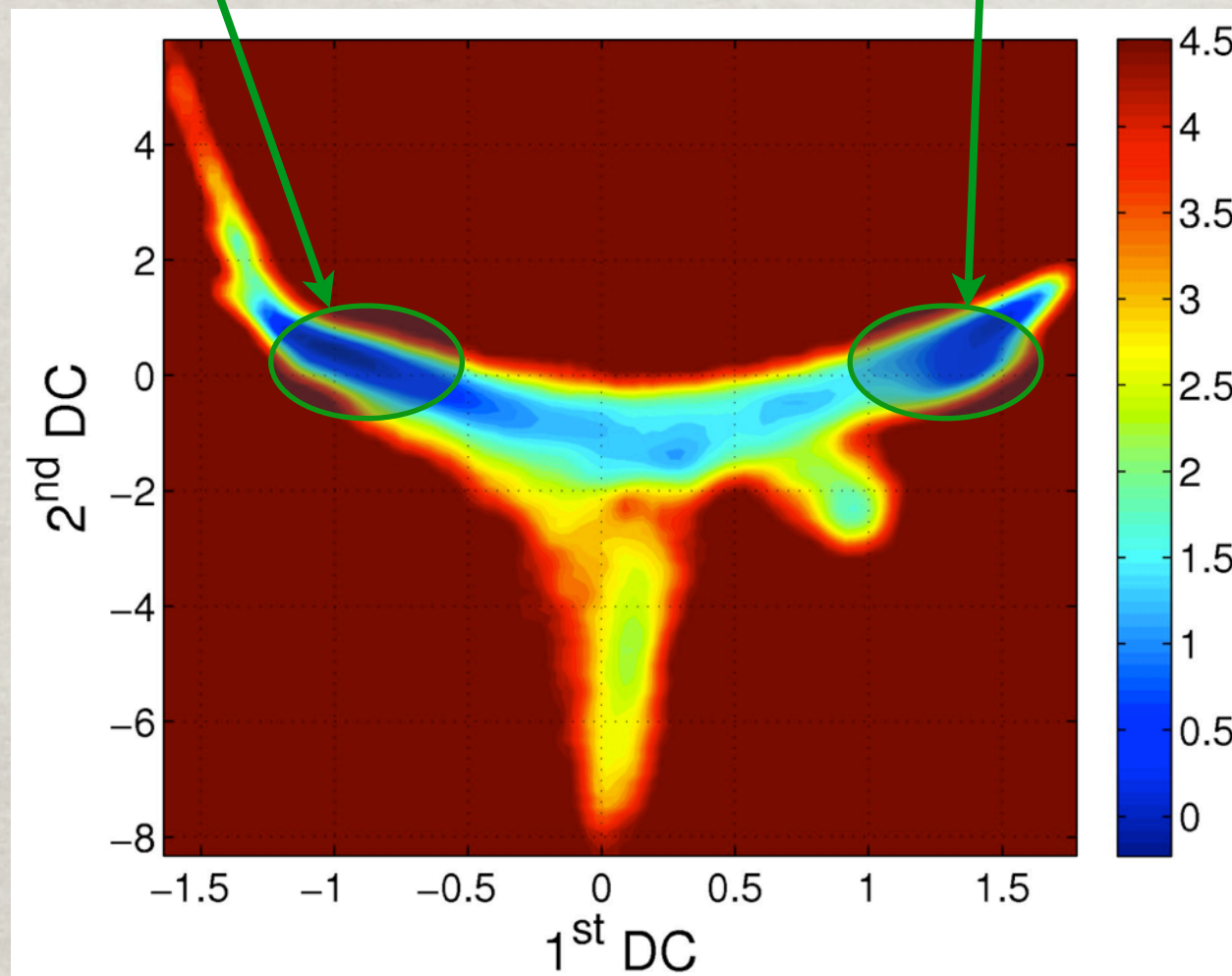
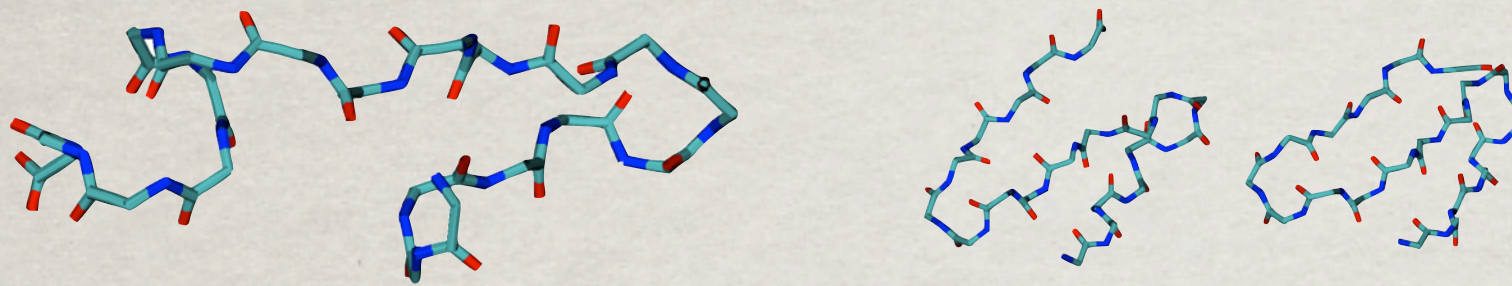


Qi B., Muff S., Caflisch A., Dinner A.,  
J. Phys. Chem. B 2010, 114, 6979



# Example: b3s

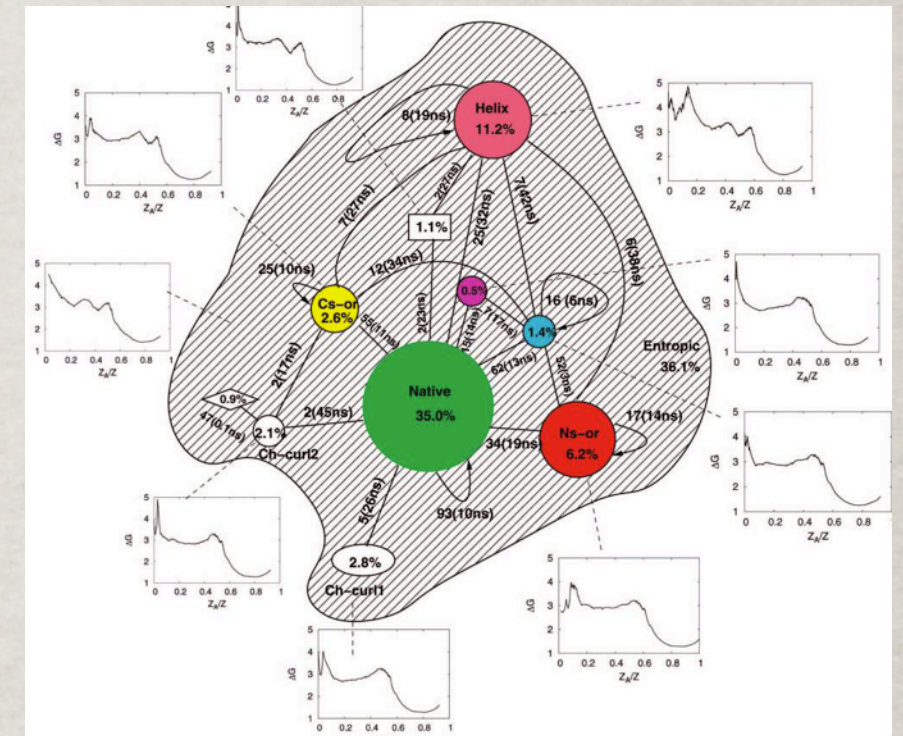
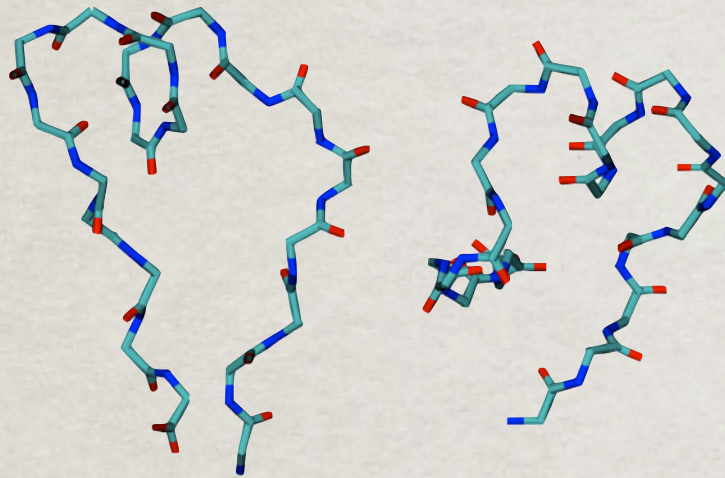
The main folding process



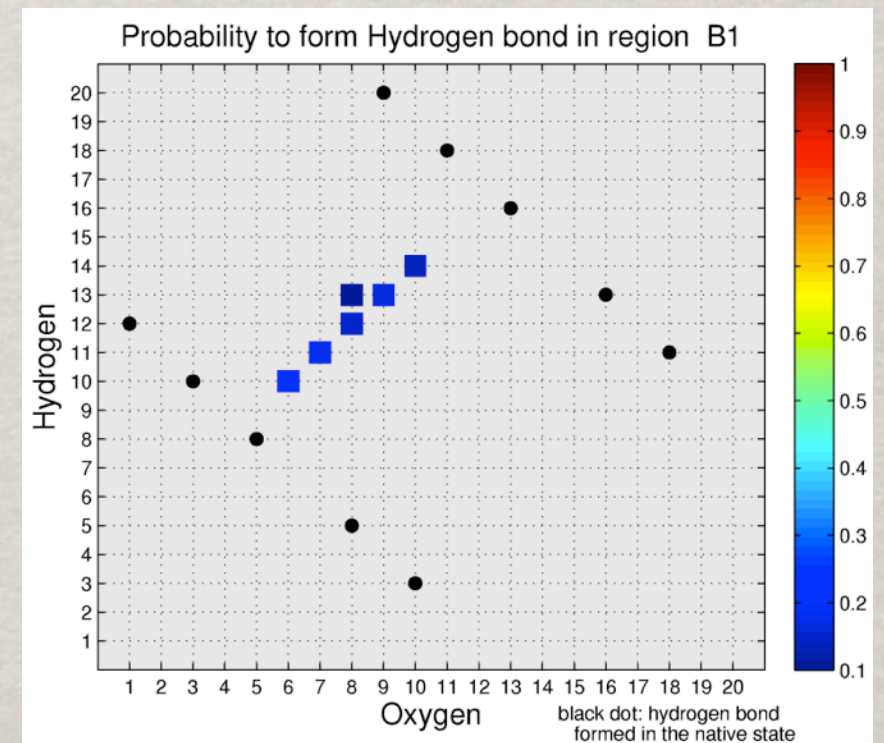
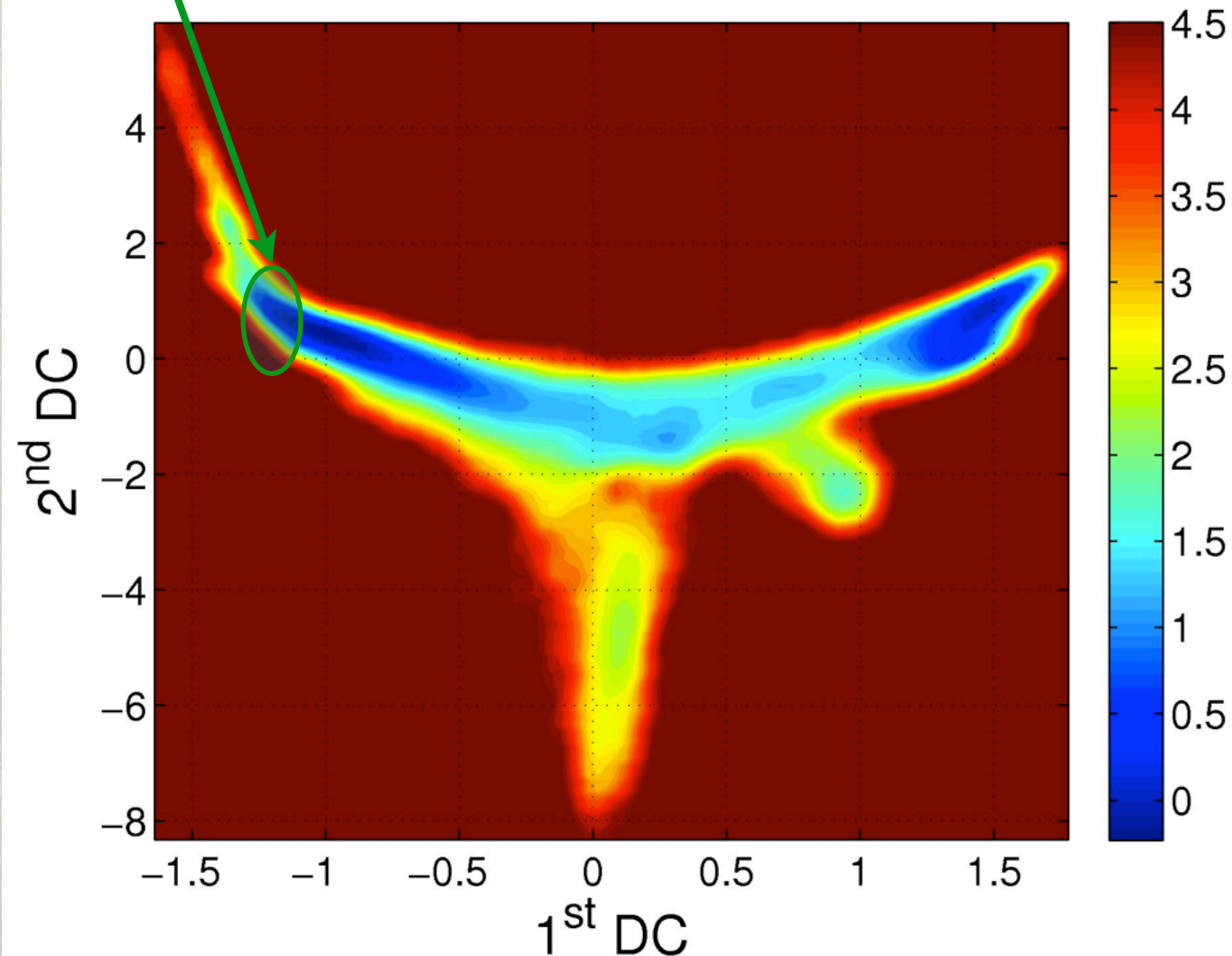


# Example: b3s

The helix state



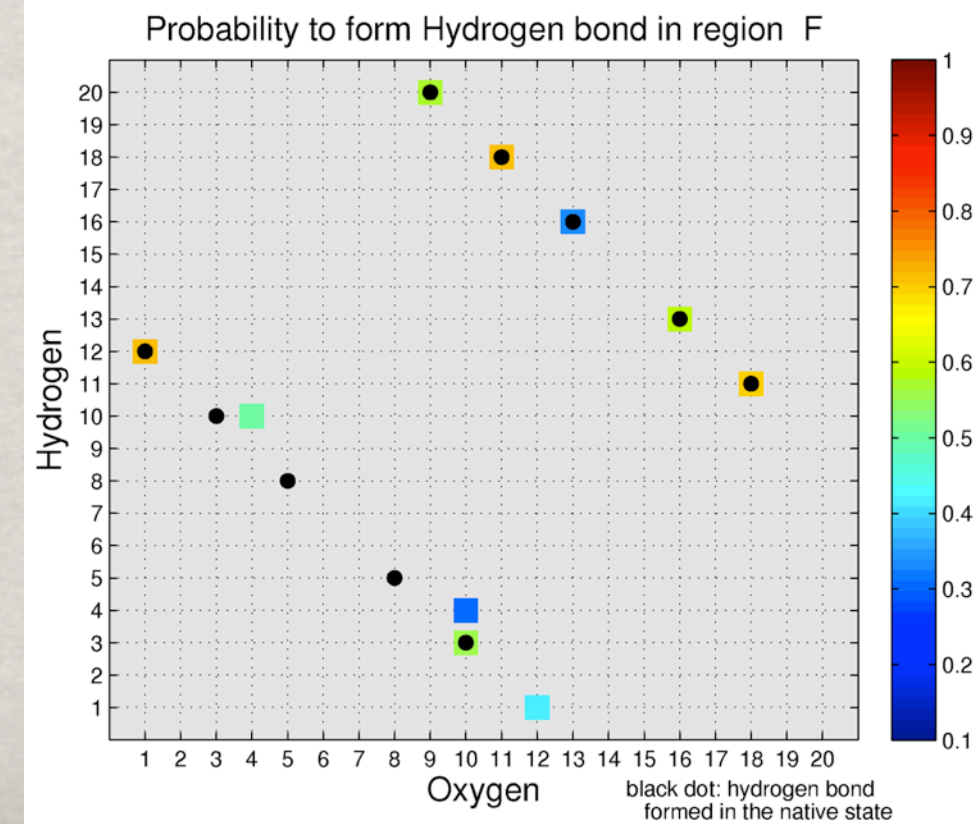
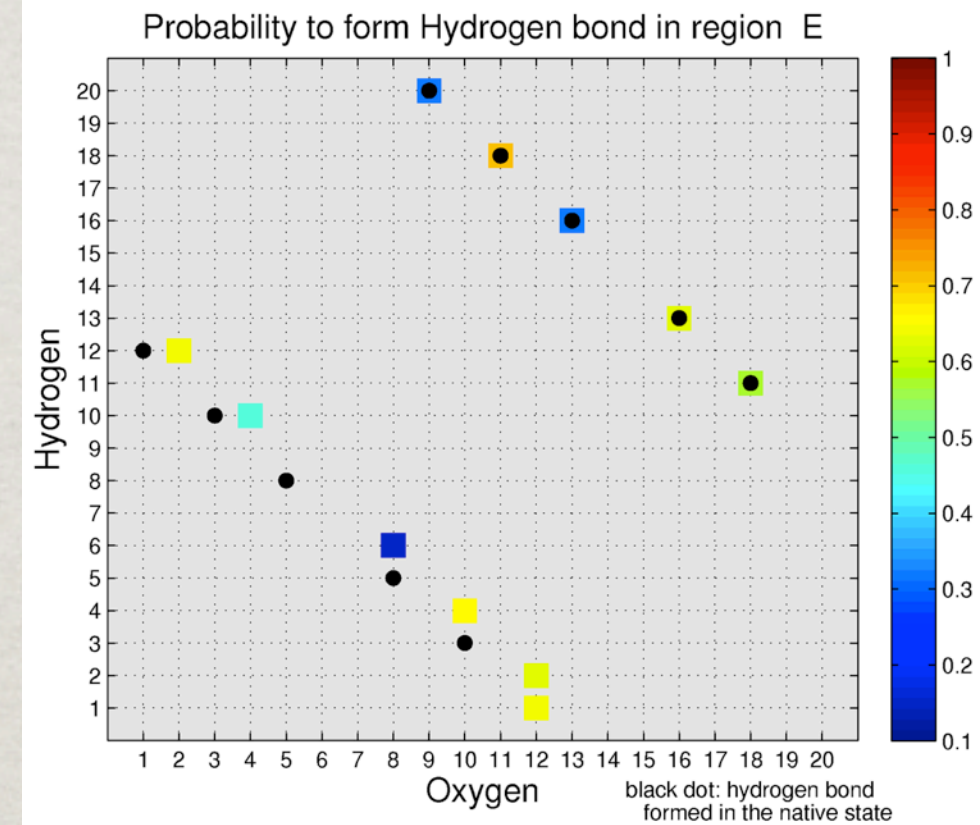
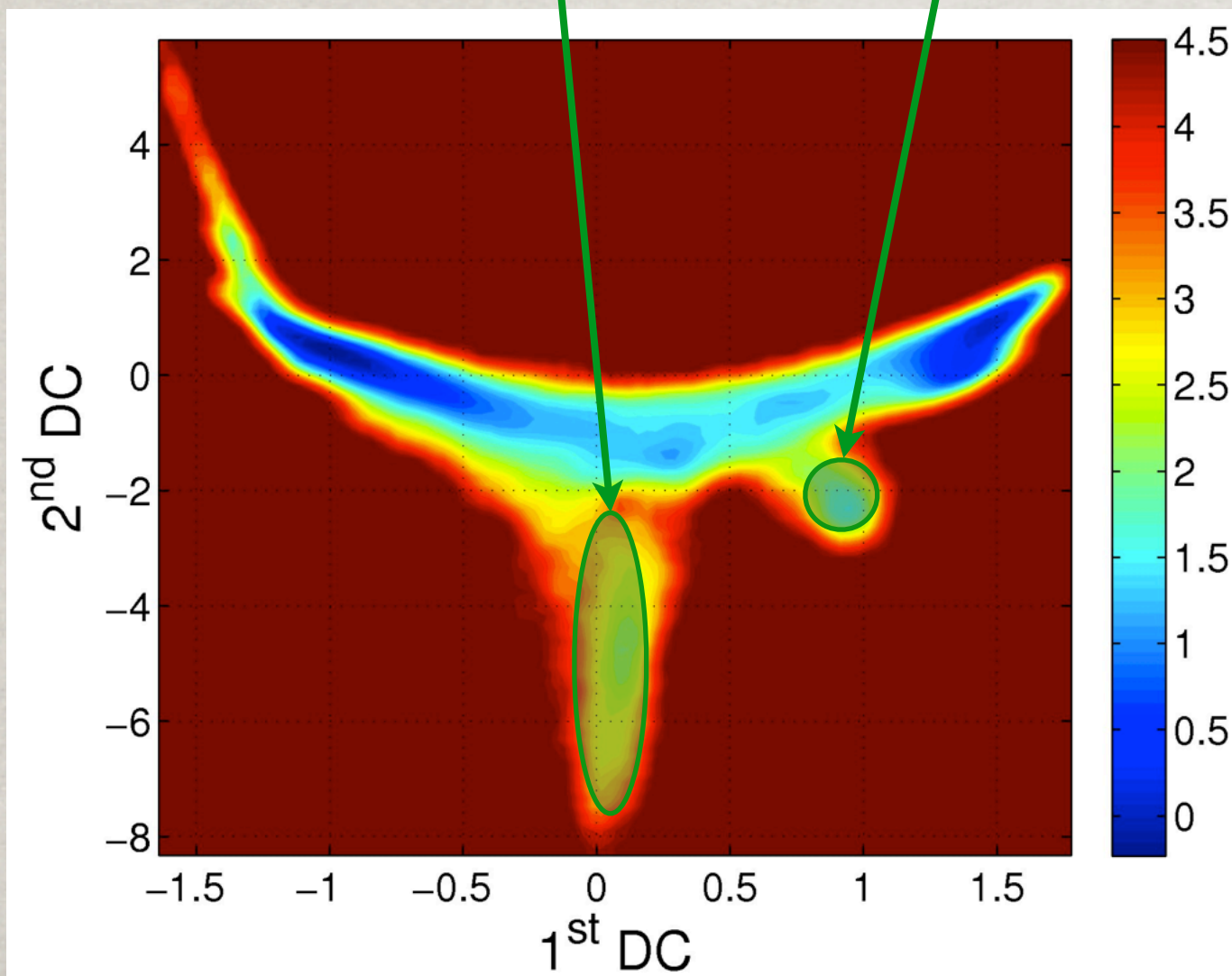
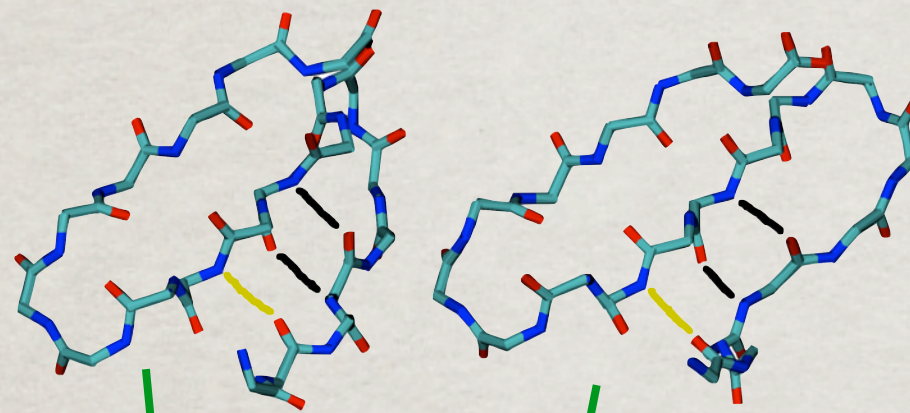
Qi B., Muff S., Caflisch A., Dinner A.,  
J. Phys. Chem. B 2010, 114, 6979





# Example: b3s

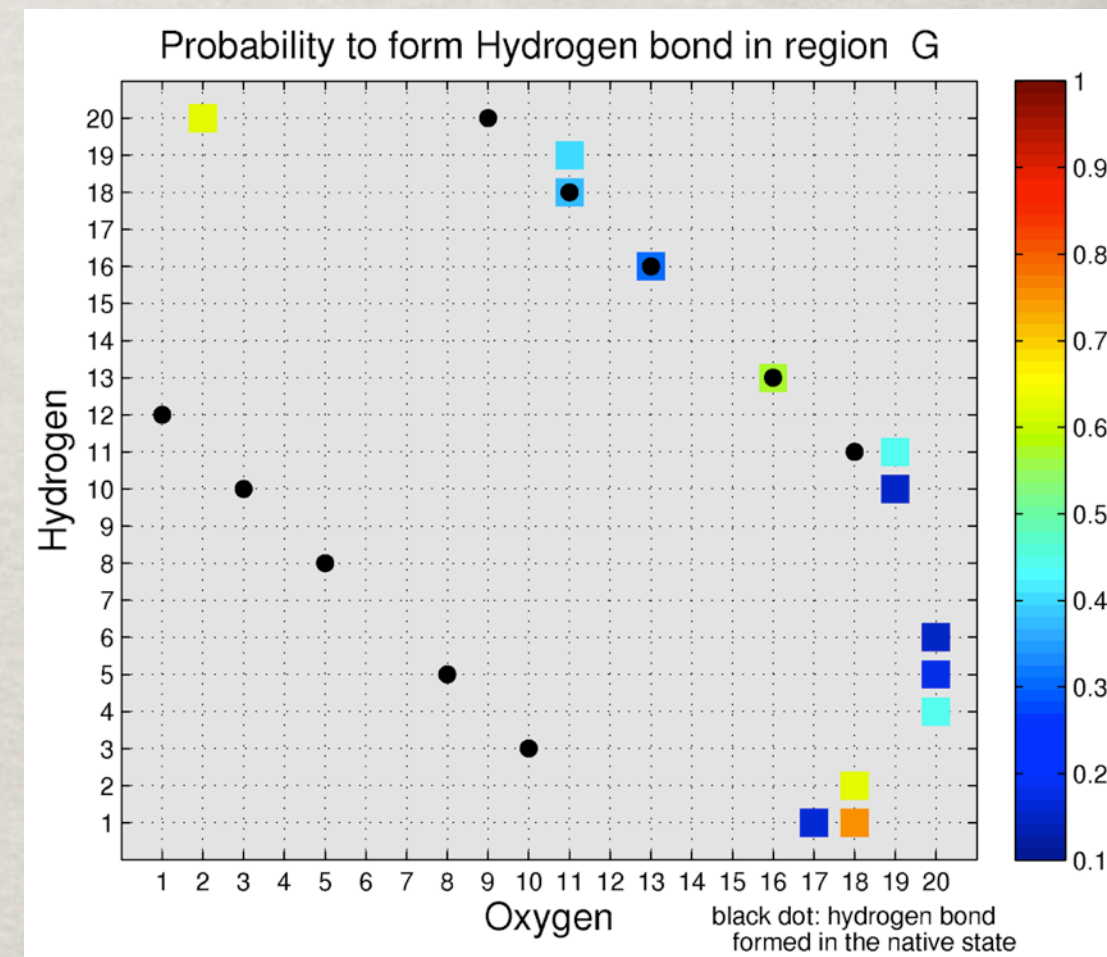
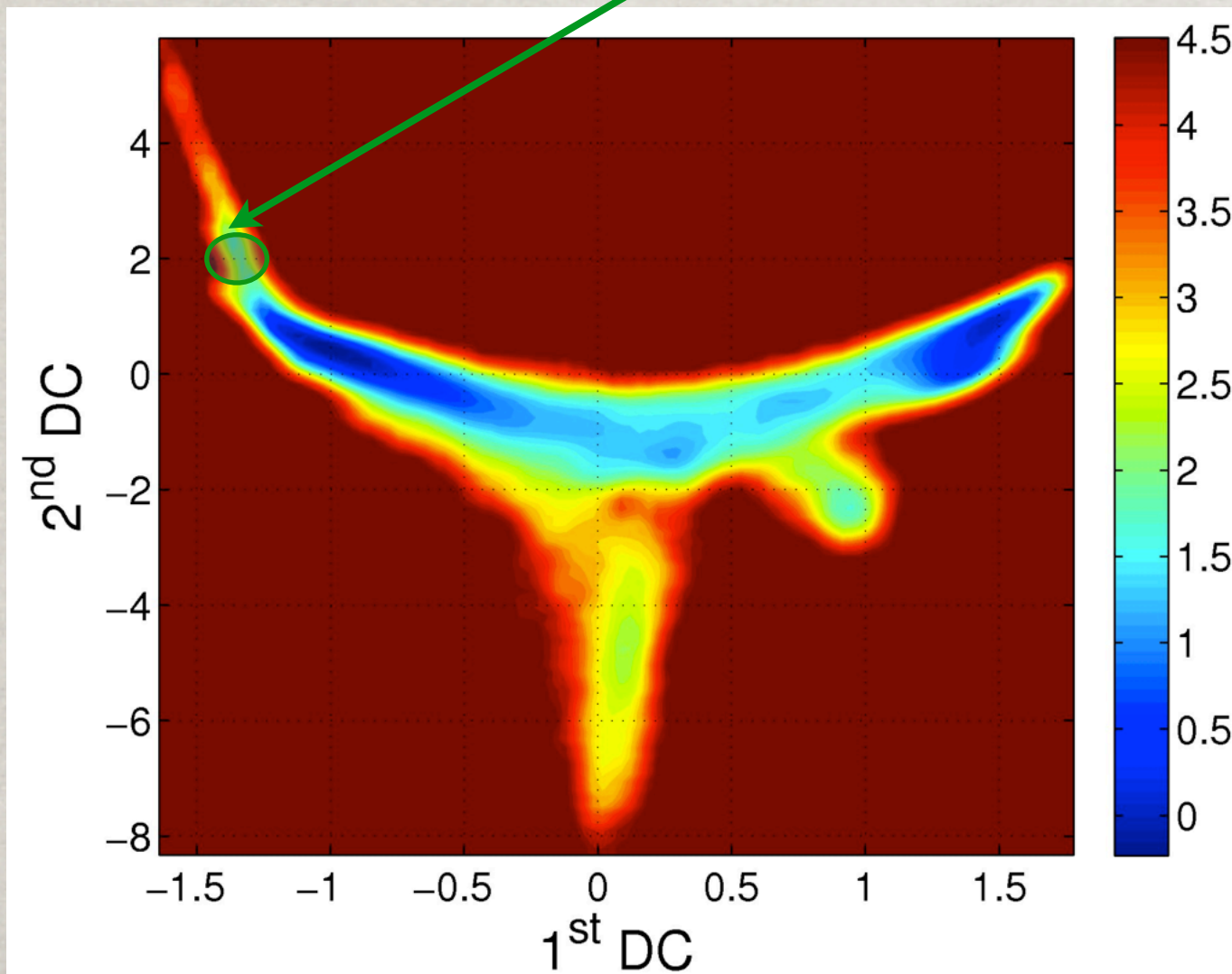
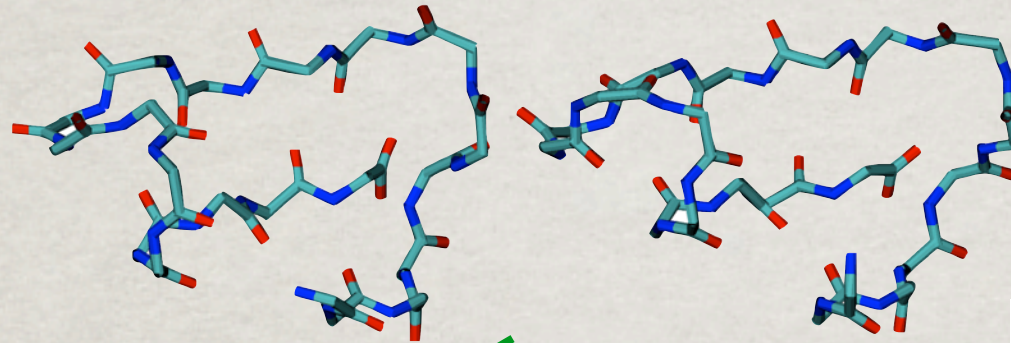
Misfolded states





# Example: b3s

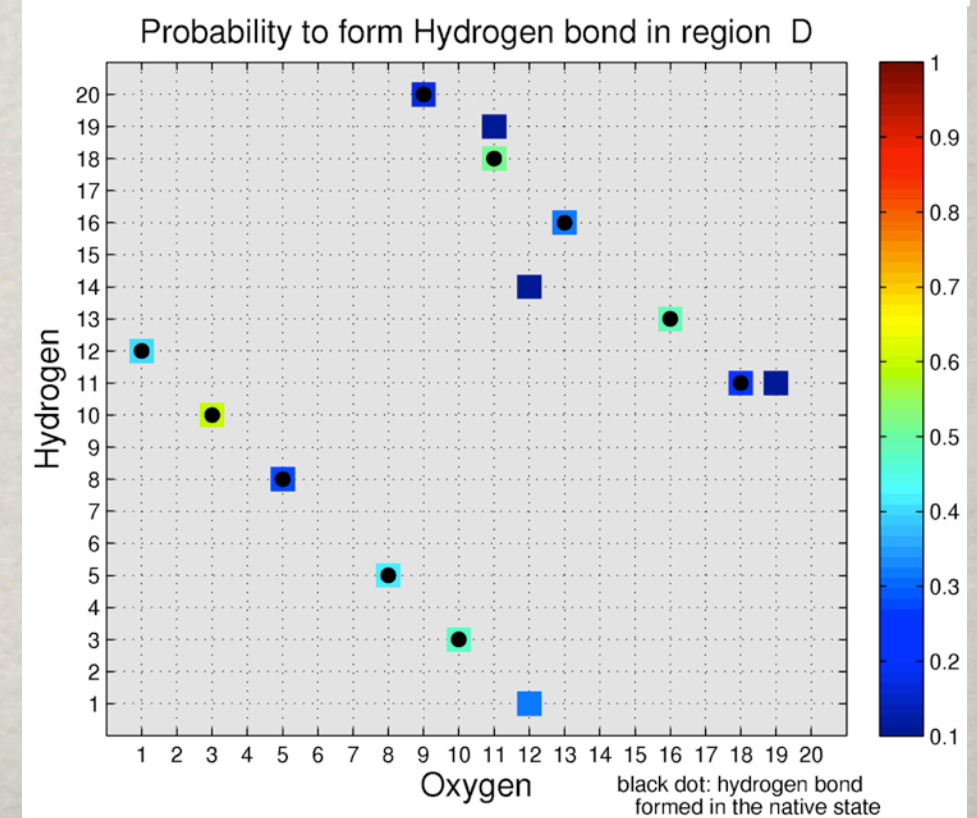
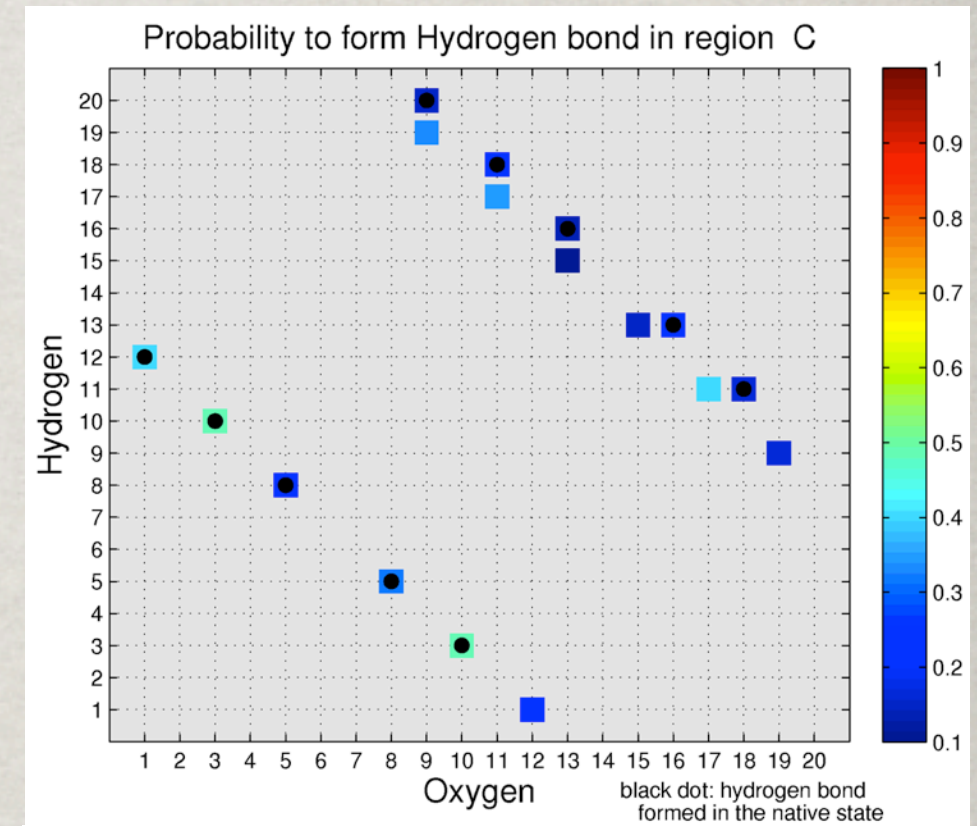
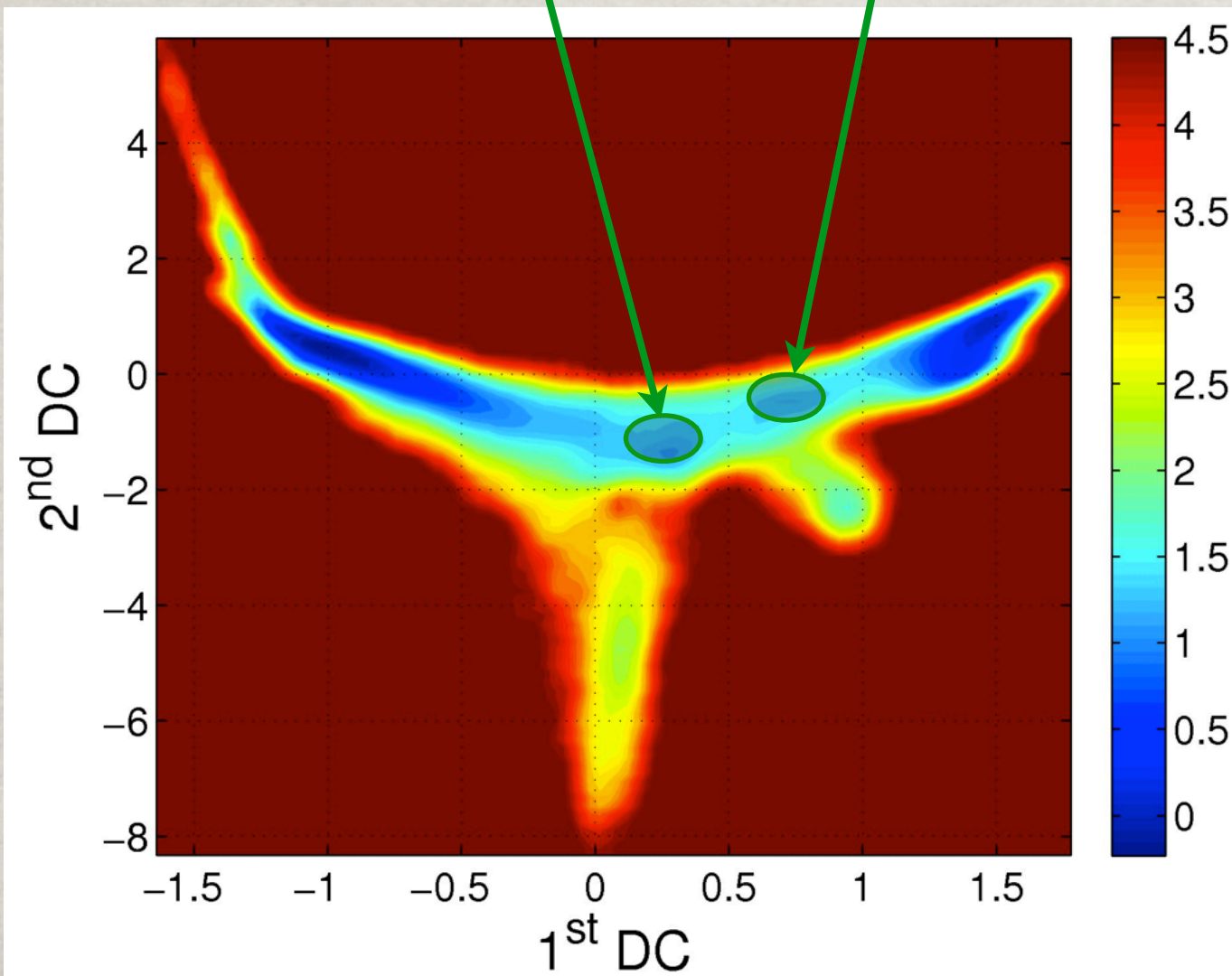
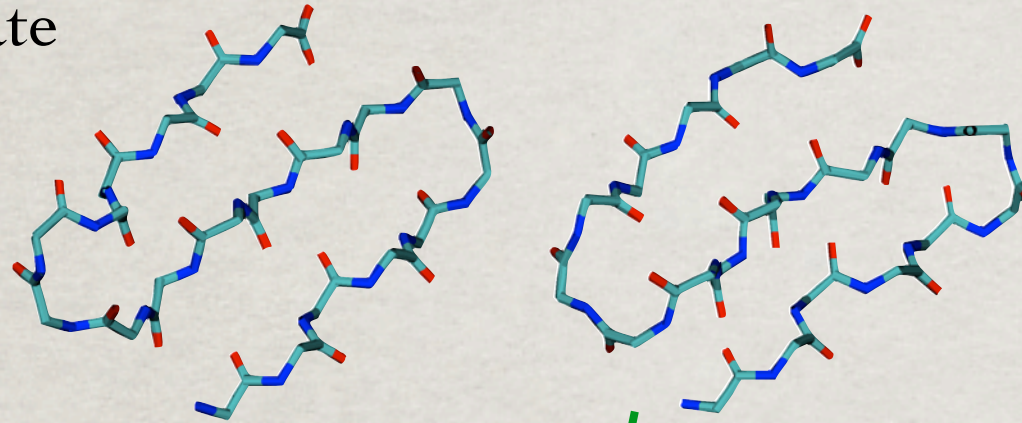
Another  
misfolded state





# Example: b3s

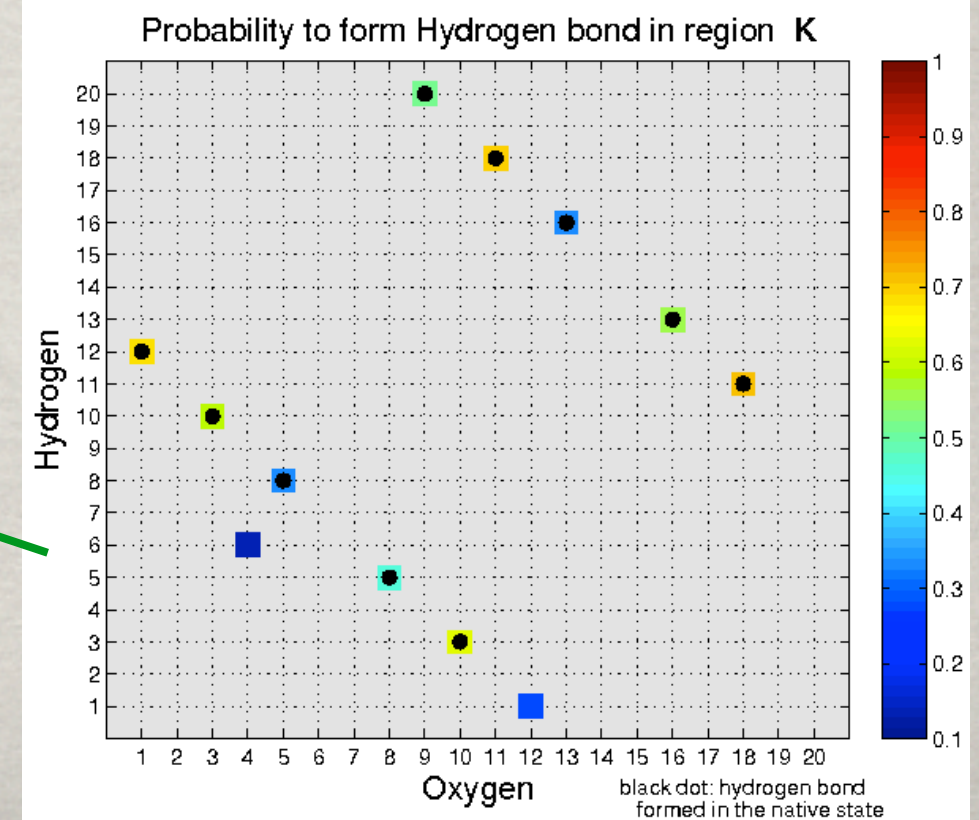
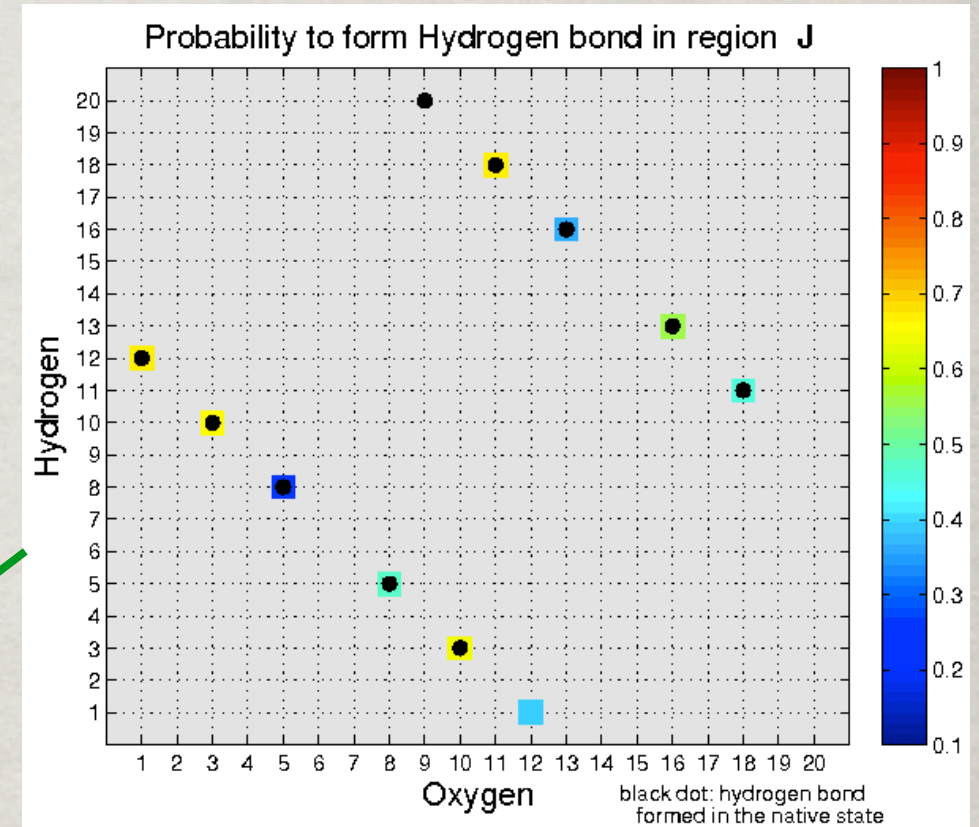
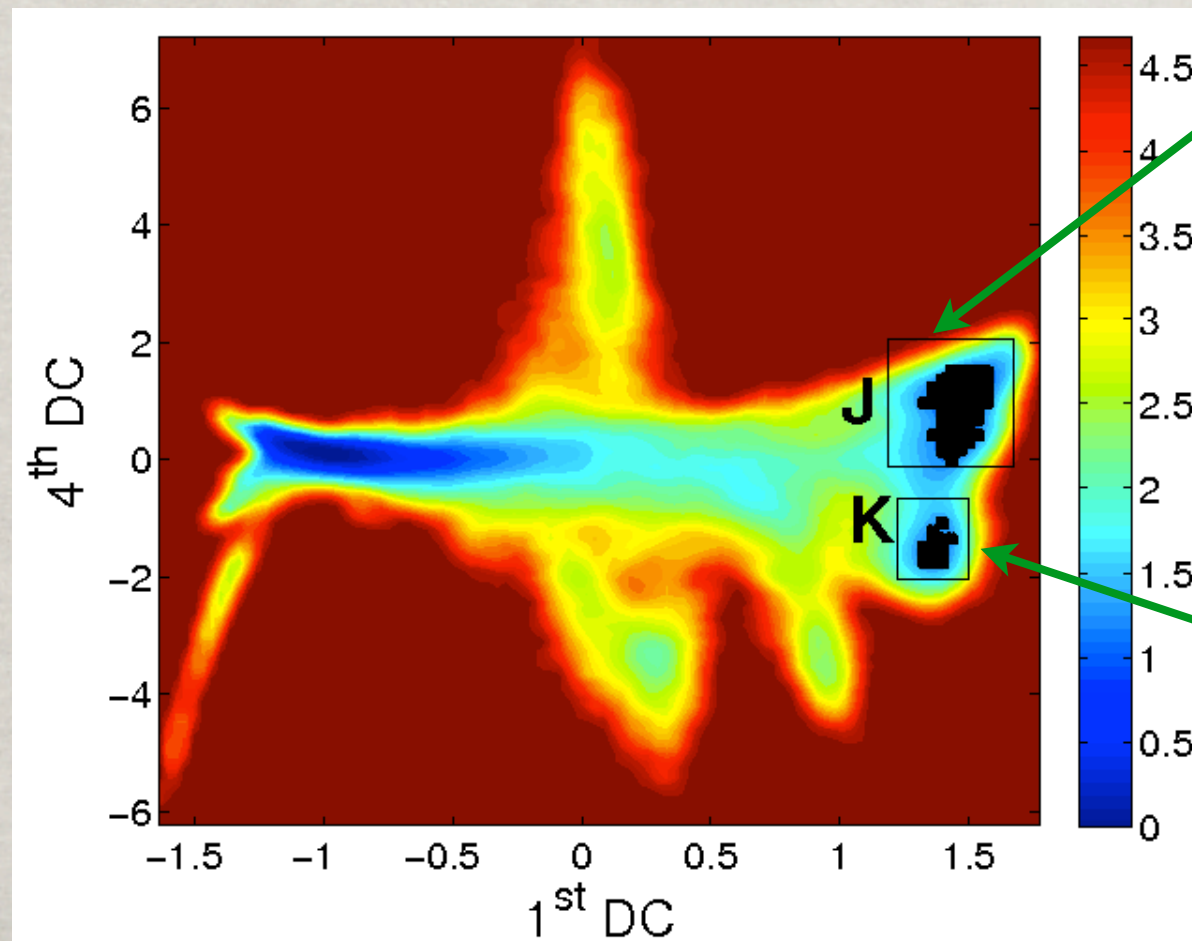
The intermediate  
state





# Example: b3s

The 4th DC shows the folded state splits into two sub-states. Sub-states J and K correspond to fluctuations in the native state. Compared to the completely folded structure, in state J the last hydrogen bond in the C-terminus  $\beta$ -sheet is not formed; in state K all the hydrogen bonds are formed and an additional one can be formed, although with low probability.





# Future directions/other projects

- . More general sampling schemes
- . Exploiting geometry for adaptive sampling
- . Nonlinear versions of geometric properties
- . Larger proteins, with higher dimensional dynamics
- . Analysis of time series of graphs and points clouds, and related multiscale distances.
- . Multiscale homogenization of random walks on graphs.
- . Active learning and visualization of large data sets.

Collaborators: D. Brady (EE, Duke), R. Brady (CS, Duke), C. Clementi (Chem, Rice), J. Mattingly (Math, Duke), E. Monson (CS, Duke), S. Mukherjee (Stats, Duke), R. Rajae (EE, Oregon), M. Rohrdanz (Rice), R. Schul (Math, Stony Brook), W. Willinger (AT&T)

THANK YOU!

[www.math.duke.edu/~mauro](http://www.math.duke.edu/~mauro)

