

# A Geometric Perspective on Machine Learning

**Partha Niyogi**

**The University of Chicago**

Collaborators: M. Belkin, V. Sindhvani, X. He, S. Smale, S. Weinberger

# Manifold Learning

Learning when data  $\sim \mathcal{M} \subset \mathbb{R}^N$

- Clustering:  $\mathcal{M} \rightarrow \{1, \dots, k\}$

connected components, min cut

- Classification:  $\mathcal{M} \rightarrow \{-1, +1\}$

$P$  on  $\mathcal{M} \times \{-1, +1\}$

- Dimensionality Reduction:  $f : \mathcal{M} \rightarrow \mathbb{R}^n \quad n \ll N$

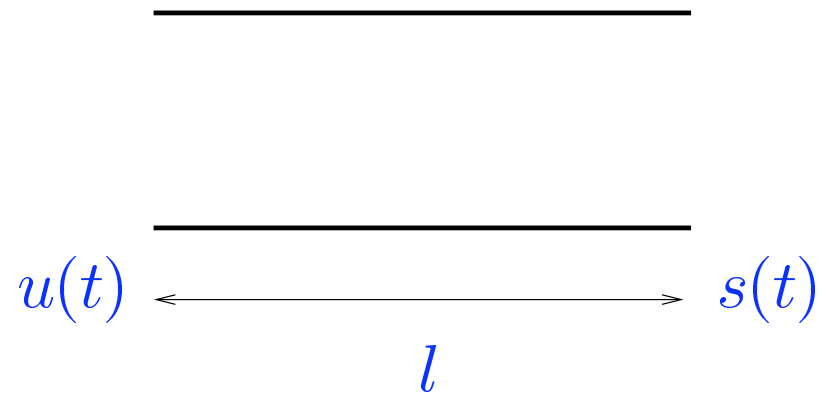
- $\mathcal{M}$  unknown: what can you learn about  $\mathcal{M}$  from data?

e.g. dimensionality, connected components

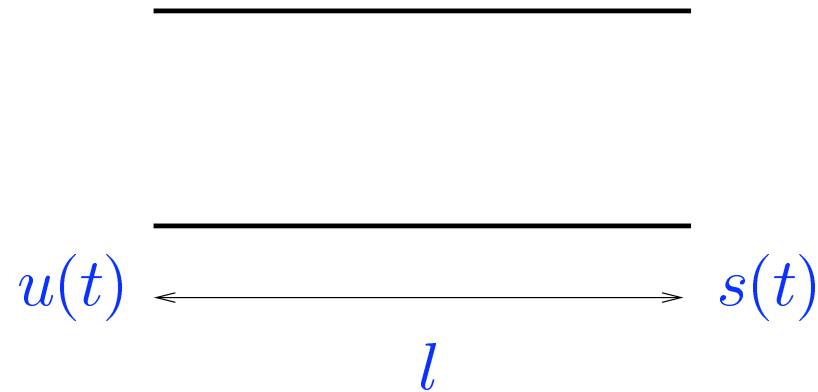
holes, handles, homology

curvature, geodesics

# An Acoustic Example



# An Acoustic Example



One Dimensional Air Flow

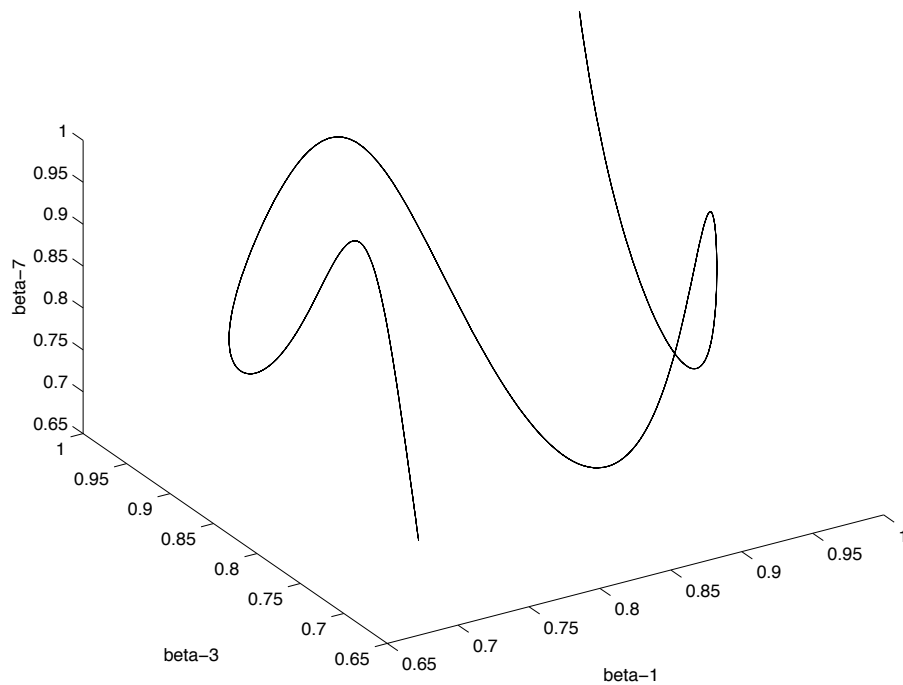
$$(i) \frac{\partial V}{\partial x} = -\frac{A}{\rho c^2} \frac{\partial P}{\partial t}$$

$$(ii) \frac{\partial P}{\partial x} = -\frac{\rho}{A} \frac{\partial V}{\partial t}$$

$V(x, t)$  = volume velocity

$P(x, t)$  = pressure

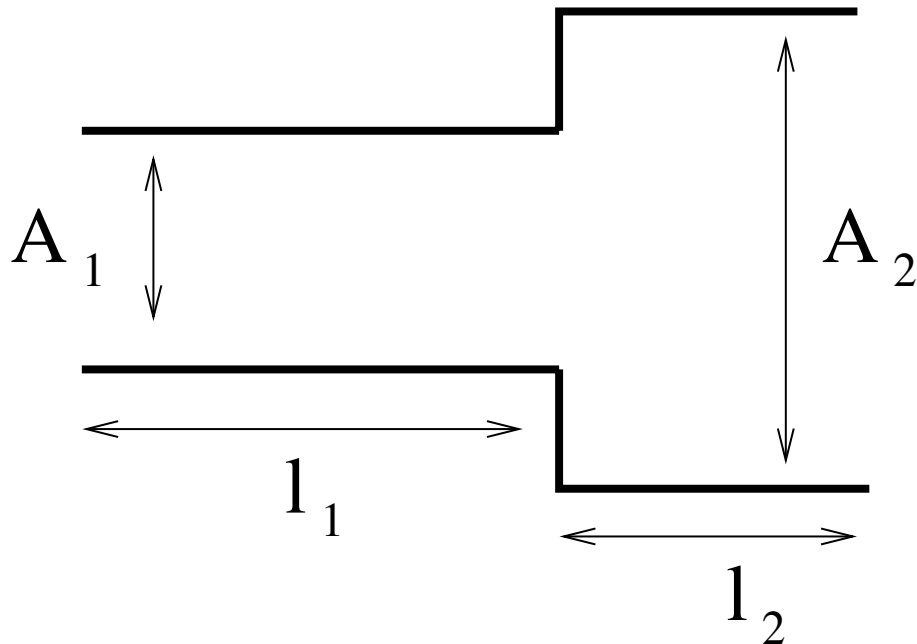
# Solutions



$$u(t) = \sum_{n=1}^{\infty} \alpha_n \sin(n\omega_0 t) \in l_2$$

$$s(t) = \sum_{n=1}^{\infty} \beta_n \sin(n\omega_0 t) \in l_2$$

# Acoustic Phonetics



Vocal Tract modeled as a sequence of tubes.  
(e.g. Stevens, 1998)

Jansen and Niyogi (in prep.)

# Pattern Recognition

$P$  on  $X \times Y$

$$X = \mathbb{R}^N; Y = \{0, 1\}, \mathbb{R}$$

$(x_i, y_i)$  labeled examples

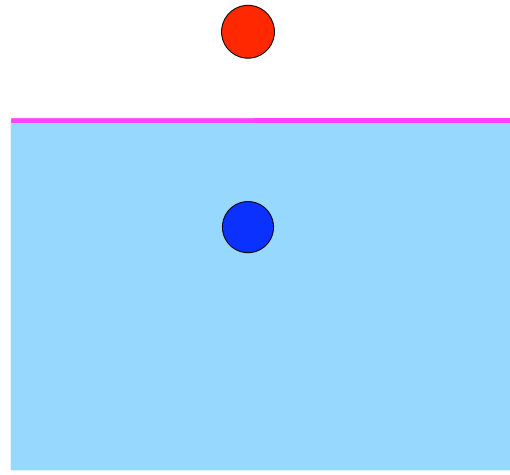
find  $f : X \rightarrow Y$

*Ill Posed*

# Simplicity



# Simplicity



# Regularization Principle

$$f = \arg \min_{f \in H_K} \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2 + \gamma \|f\|_K^2$$

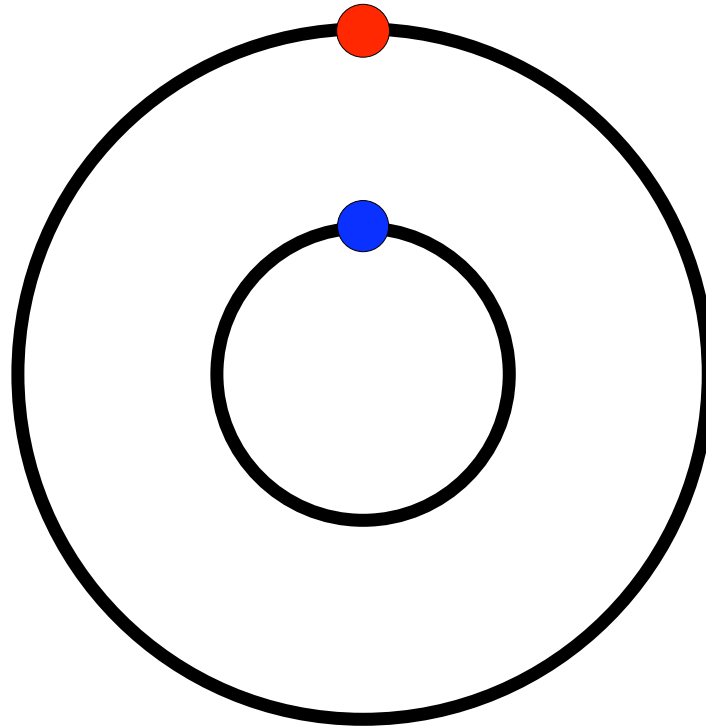
Splines

Ridge Regression

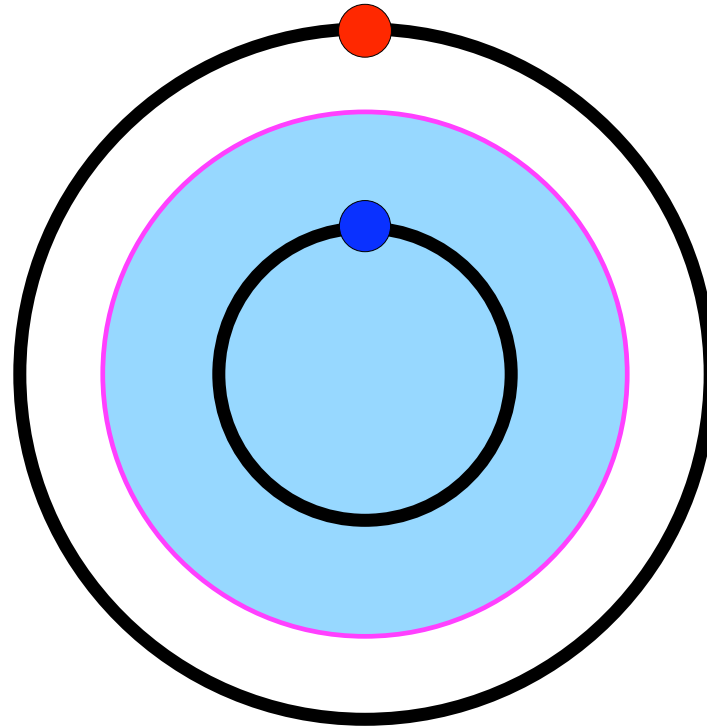
SVM

- $K : X \times X \rightarrow \mathbb{R}$  is a p.d. kernel  
e.g.  $e^{-\frac{\|x-y\|^2}{\sigma^2}}$ ,  $(1 + x \cdot y)^d$ , etc.
- $H_K$  is a corresponding RKHS  
e.g., certain *Sobolev* spaces, polynomial families, etc.

# Simplicity is Relative



# Simplicity is Relative



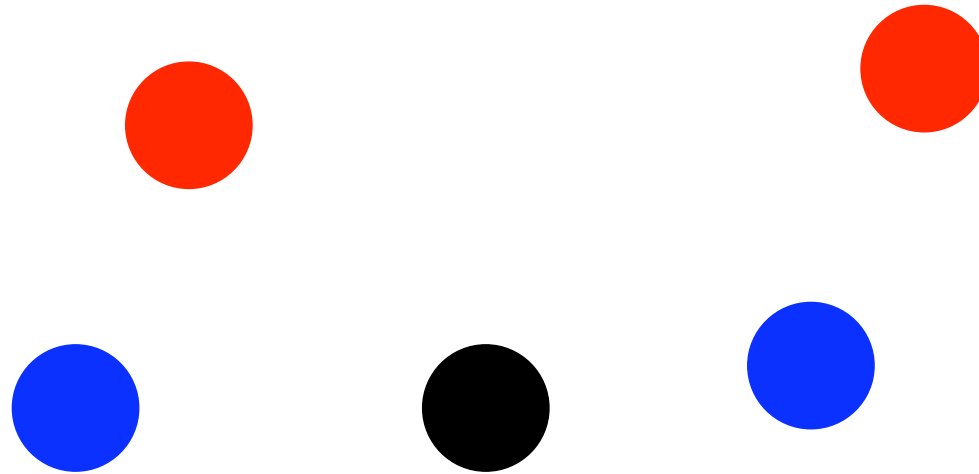
# Intuitions

- $\text{supp } P_X$  has manifold structure
- *geodesic* distance v/s *ambient* distance
- geometric structure of data should be incorporated
- $f$  versus  $f_{\mathcal{M}}$

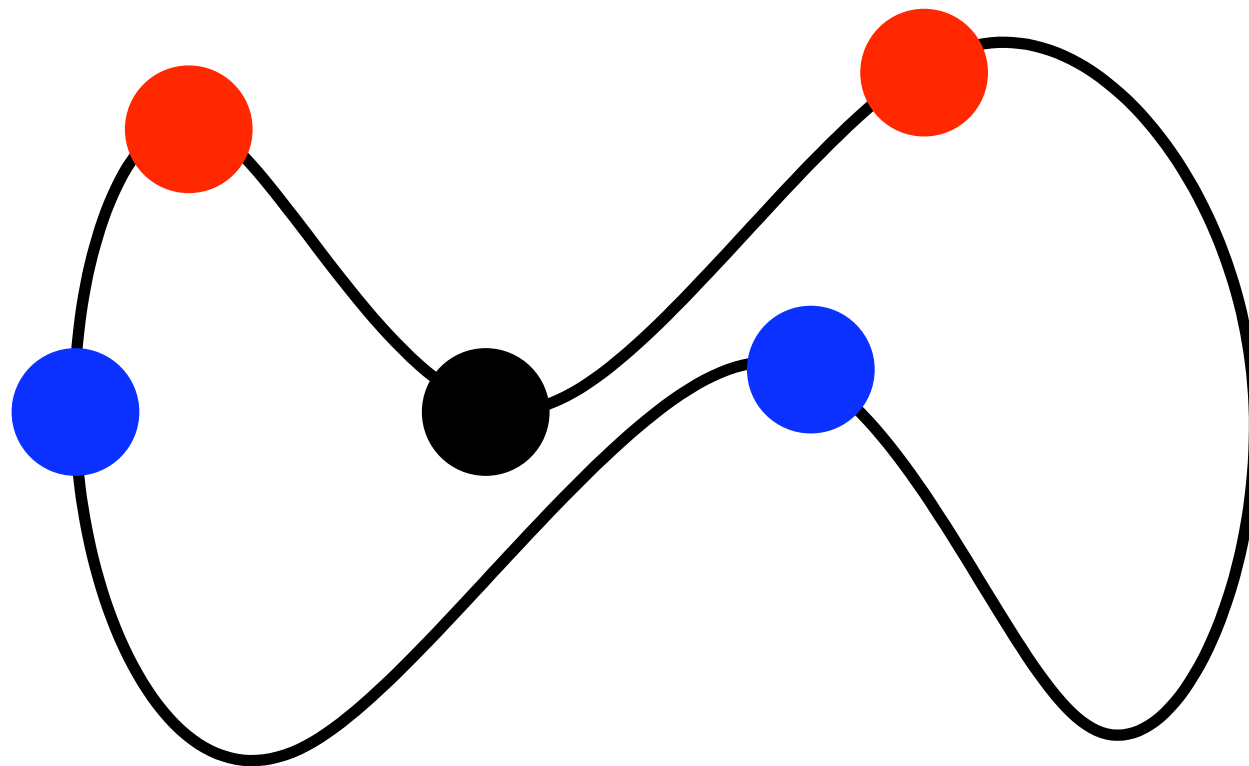
# Geometry and similarity



# Geometry and similarity

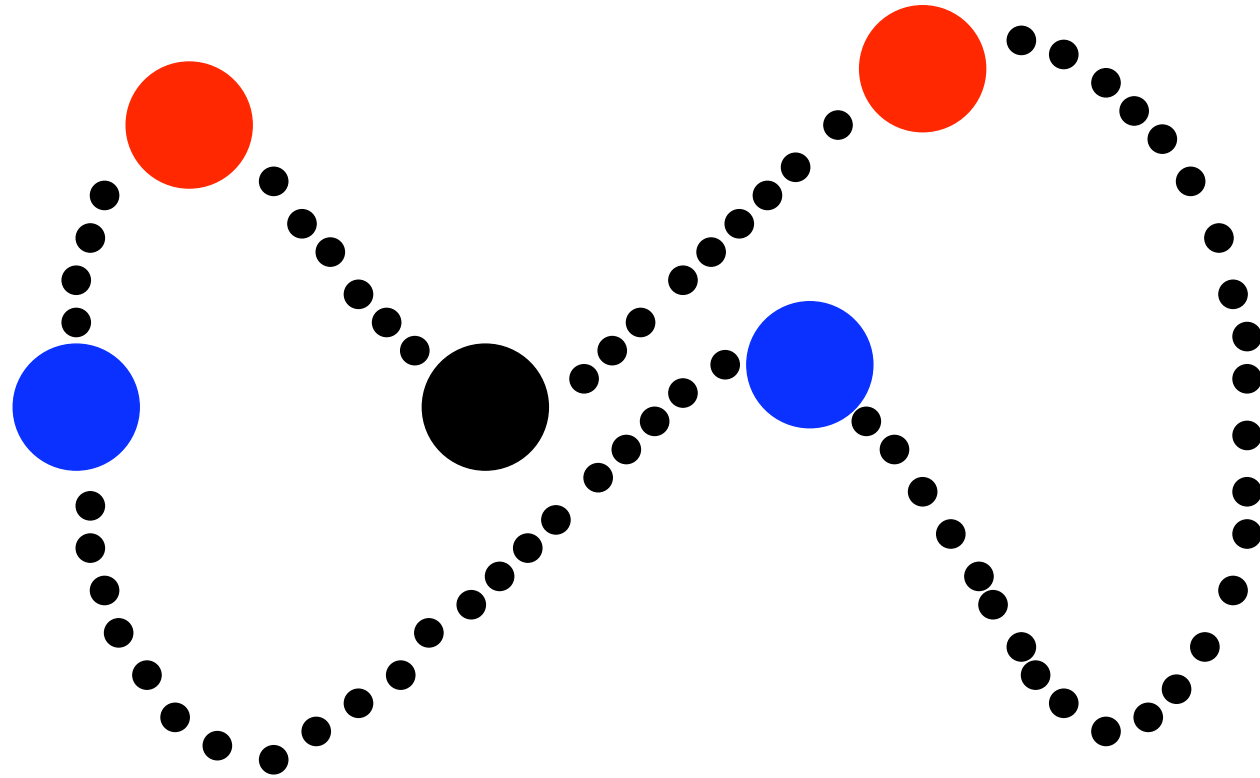


# Geometry and similarity



Geometry is important.

# Geometry and similarity



Geometry is important.  
Unlabeled data to estimate geometry.

# Manifold Regularization

$$\min_{f \in H_K} \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2 + \gamma_A \|f\|_K^2 + \gamma_I \|f\|_I^2$$

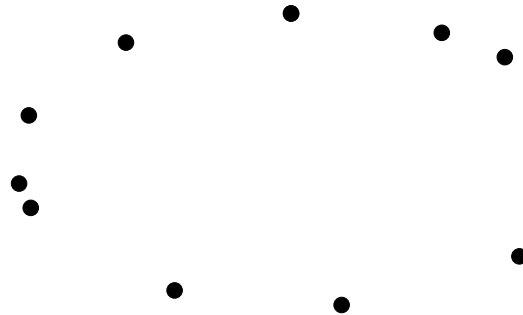
$$\|f\|_I^2 = \begin{cases} \text{Laplacian} & \int \langle \text{grad}_{\mathcal{M}} f, \text{grad}_{\mathcal{M}} f \rangle = \int f \Delta_{\mathcal{M}} f \\ \text{Iterated Laplacian} & \int f \Delta_{\mathcal{M}}^i f \\ \text{Heat kernel} & e^{-\Delta_{\mathcal{M}} t} \\ \text{Differential Operator} & \int f(Df) \end{cases}$$

Representer Theorem:  $f = \sum_{i=1}^n \alpha_i K(x, x_i) + \int_{\mathcal{M}} \alpha(y) K(x, y)$

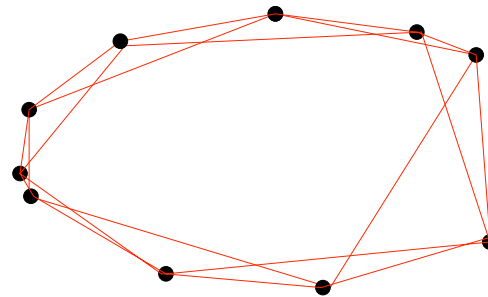
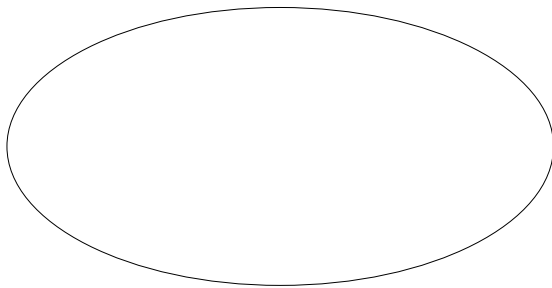
Belkin, Niyogi, Sindhwani (2004)

# Approximating $\|f\|_I^2$

$\mathcal{M}$  is unknown but  $x_1 \dots x_M \in \mathcal{M}$



$$\|f\|_I^2 = \int_{\mathcal{M}} \langle \nabla_{\mathcal{M}} f, \nabla_{\mathcal{M}} f \rangle \approx \sum_{i \sim j} W_{ij} (f(x_i) - f(x_j))^2$$



# Manifolds and Graphs

$$\mathcal{M} \approx G = (V, E)$$

$$e_{ij} \in E \text{ if } \|x_i - x_j\| < \epsilon$$

$$W_{ij} = e^{-\frac{\|x_i - x_j\|^2}{t}}$$

$$\Delta_{\mathcal{M}} \approx L = D - W$$

$$\int \langle \text{grad } f, \text{grad } f \rangle \approx \sum_{i,j} W_{ij} (f(x_i) - f(x_j))^2$$

$$\int f(\Delta f) \approx \mathbf{f}^T L \mathbf{f}$$

# Manifold Regularization

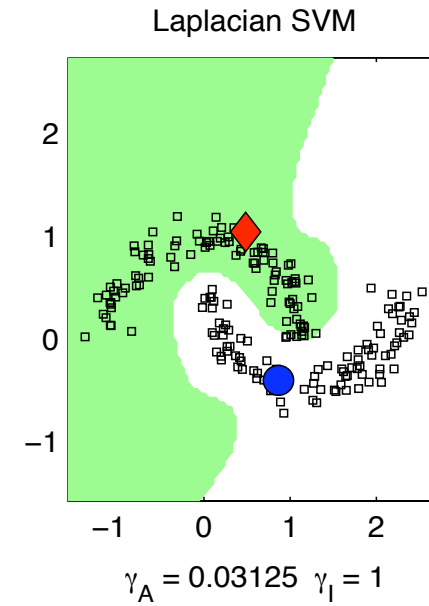
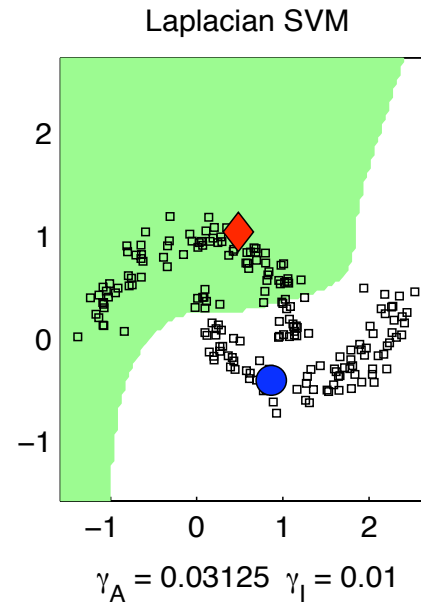
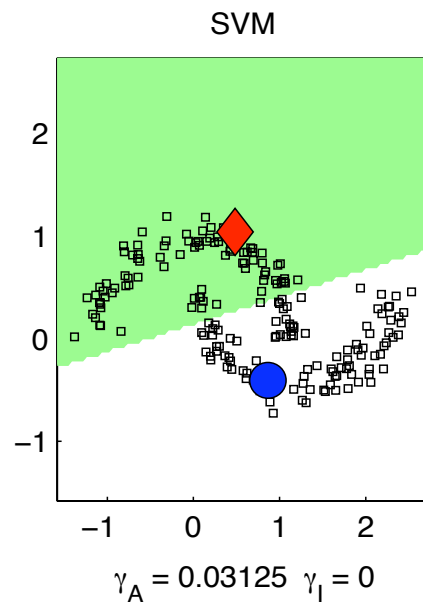
$$\frac{1}{n} \sum_{i=1}^n V(f(x_i), y_i) + \gamma_A \|f\|_K^2 + \gamma_I \sum_{i \sim j} W_{ij} (f(x_i) - f(x_j))^2$$

Representer Theorem:  $f_{opt} = \sum_{i=1}^{n+m} \alpha_i K(x, x_i)$

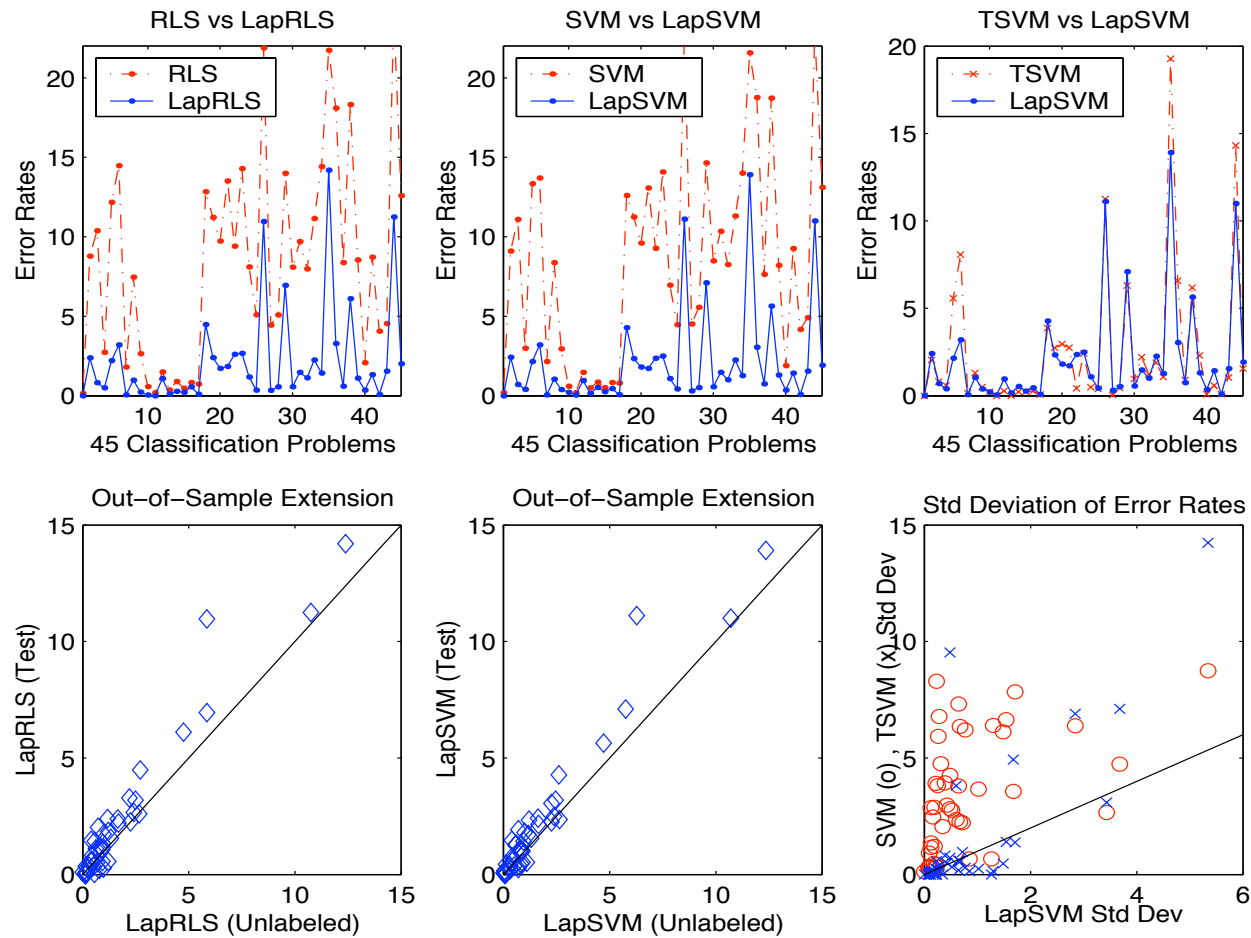
$V(f(x), y) = (f(x) - y)^2$ : Least squares

$V(f(x), y) = (1 - yf(x))_+$ : Hinge loss (Support Vector Machines)

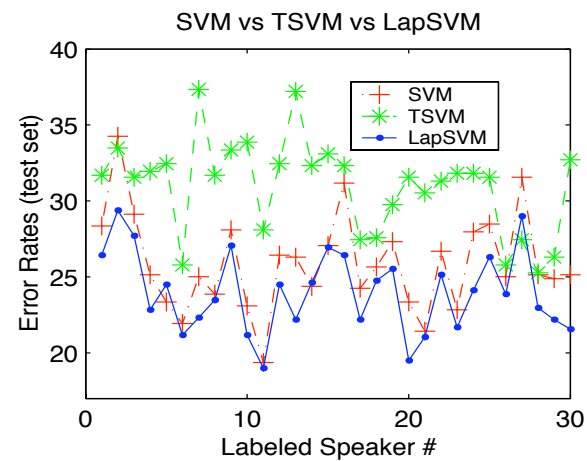
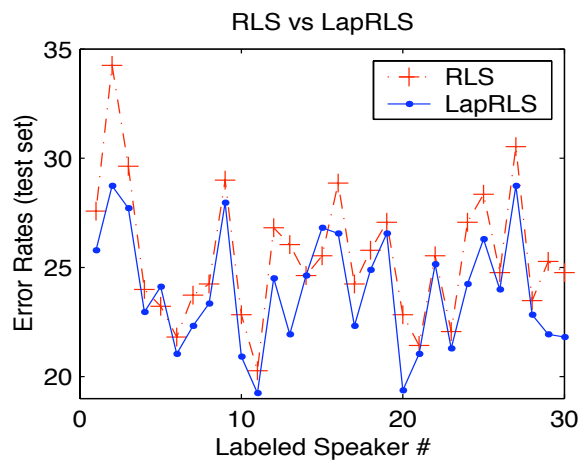
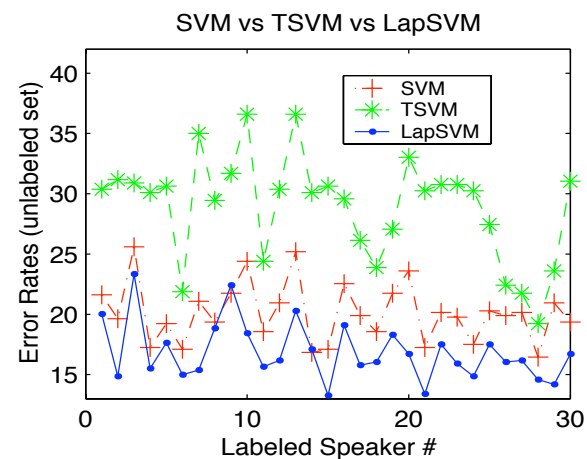
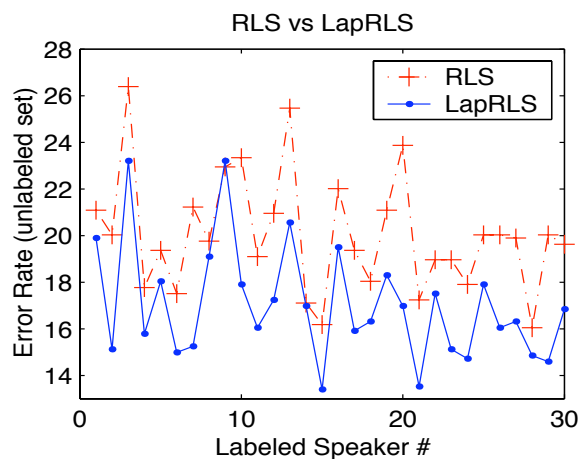
# Ambient and Intrinsic Regularization



# Experimental Results: USPS



# Experimental Results: Isolet



# Experimental comparisons

Dataset → Algorithm ↓	g50c	Coil20	Uspst	mac-win	WebKB (link)	WebKB (page)	WebKB (page+link)
SVM (full labels)	3.82	0.0	3.35	2.32	6.3	6.5	1.0
RLS (full labels)	3.82	0.0	2.49	2.21	5.6	6.0	2.2
SVM (l labels)	8.32	24.64	23.18	18.87	25.6	22.2	15.6
RLS (l labels)	8.28	25.39	22.90	18.81	28.0	28.4	21.7
Graph-Reg	17.30	6.20	21.30	11.71	22.0	10.7	6.6
TSVM	6.87	26.26	26.46	7.44	14.5	8.6	7.8
Graph-density	8.32	6.43	16.92	10.48	-	-	-
$\nabla$ TSVM	5.80	17.56	17.61	5.71	-	-	-
LDS	5.62	4.86	15.79	5.13	-	-	-
LapSVM	<b>5.44</b>	<b>3.66</b>	<b>12.67</b>	10.41	18.1	10.5	<b>6.4</b>
LapRLS	<b>5.18</b>	<b>3.36</b>	<b>12.69</b>	10.01	19.2	11.0	6.9
LapSVM <sub>joint</sub>	-	-	-	-	<b>5.7</b>	<b>6.7</b>	<b>6.4</b>
LapRLS <sub>joint</sub>	-	-	-	-	<b>5.6</b>	<b>8.0</b>	<b>5.8</b>

# Convergence

$$E_{opt} = \min_{f \in H_K} E[(y - f(x))^2]$$

$$\uparrow \gamma_A, \gamma_I \rightarrow 0$$

$$E_{\gamma_A, \gamma_I} = \min_{f \in H_K} E[(y - f(x))^2] + \gamma_A \|f\|_K^2 + \gamma_I \int f \Delta_{\mathcal{M}} f$$

$$\uparrow n \rightarrow \infty$$

$$E_{\gamma_A, \gamma_I, n} = \min_{f \in H_K} \frac{1}{n} \sum_{i=1}^n (y_i - f(x))^2 + \gamma_A \|f\|_K^2 + \gamma_I \int f \Delta_{\mathcal{M}} f$$

$$\uparrow m \rightarrow \infty$$

$$E_{\gamma_A, \gamma_I, n, m} = \min_{f \in H_K} \frac{1}{n} \sum_{i=1}^n (y_i - f(x))^2 + \gamma_A \|f\|_K^2 + \gamma_I \mathbf{f}^T L \mathbf{f}$$

# Convergence Theorem

with prob.  $> 1 - \delta$

$$|E_{\gamma_A, \gamma_I, n} - E_{opt}| \leq C + \gamma_A \|f_{opt}\|_K^2 + \gamma_I \int_{\mathcal{M}} f_{opt}(\Delta^l f_{opt})$$

$$C = \frac{4}{\beta^{3/2}} \sqrt{\frac{1}{n} \log\left(\frac{2}{\delta}\right)}$$

$$\beta^2 = \frac{\gamma_A}{\kappa^2} + \frac{\gamma_I}{\mu^2}$$

$$\kappa^2 = \sup_{x \in X} K(x, x)$$

$$\mu^2 = \sup_{x \in \mathcal{M}} \sum_i \left(\frac{1}{\lambda_i}\right)^l \phi_i^2(x)$$

# Graph and Manifold Laplacian

Fix  $f : X \rightarrow \mathbb{R}$ .

Fix  $x \in \mathcal{M}$

$$(L_n f) = \sum_j (f(x) - f(x_j)) e^{-\frac{\|x-x_j\|^2}{4t_n}}$$

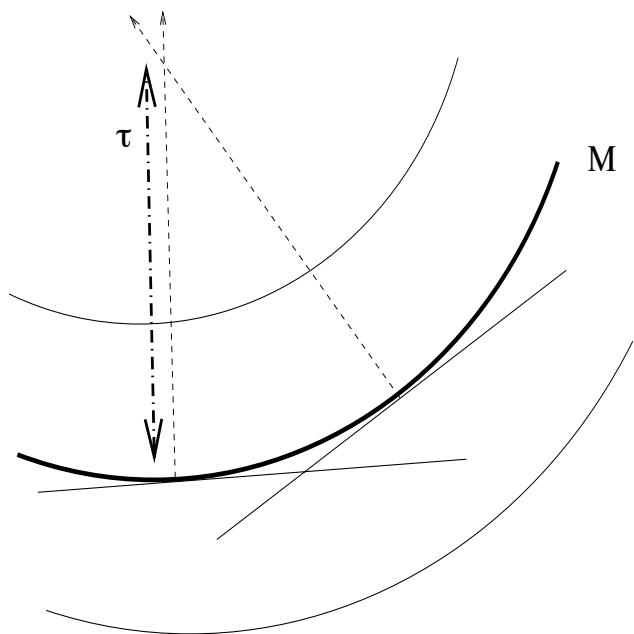
Put  $t_n = n^{-k-2-\alpha}$ , where  $\alpha > 0$

$$\text{with prob. 1, } \lim_{n \rightarrow \infty} \frac{(4\pi t_n)^{-\frac{k+1}{2}}}{n} (L_n f)|_x = \Delta_{\mathcal{M}} f|_x$$

Belkin (2003), Belkin and Niyogi (in preparation)

also Lafon (2004), Coifman et al

# Well Conditioned Submanifolds



Tubular Neighborhood

Condition No.  $\frac{1}{\tau}$

# Euclidean and Geodesic distance

$\mathcal{M} \subset \mathbb{R}^k$  condition  $\sim \tau$

$p, q \in \mathcal{M}$  where  $\|p - q\|_{\mathbb{R}^k} = d$ .

For all  $d \leq \frac{\tau}{2}$ ,

$$d_{\mathcal{M}}(p, q) \leq \tau - \tau \sqrt{1 - \frac{2d}{\tau}}$$

In fact, Second Fundamental Form Bounded by  $\frac{1}{\tau}$

# Learning Homology

$$x_1, \dots, x_n \in \mathcal{M} \subset \mathbb{R}^N$$

Can you learn **qualitative** features of  $\mathcal{M}$ ?

- Can you tell a torus from a sphere?
- Can you tell how many connected components?
- Can you tell the dimension of  $\mathcal{M}$ ?

(e.g. Carlsson, de Silva et al)

# Homology

$$x_1, \dots, x_n \in \mathcal{M} \subset \mathbb{R}^k$$

$$U = \cup_{i=1}^n B_\epsilon(x_i)$$

If  $\epsilon$  well chosen, then  $U$  deformation retracts to  $\mathcal{M}$ .

Homology of  $U$  is constructed using the *nerve* of  $U$  and agrees with the homology of  $\mathcal{M}$ .

# Theorem

$\mathcal{M} \subset \mathbb{R}^k$  with cond. no.  $\tau$

$\bar{x} = \{x_1, \dots, x_n\} \sim$  uniformly sampled i.i.d.

$$0 < \epsilon < \frac{\tau}{2} \quad \beta = \frac{\text{vol}(\mathcal{M})}{(\sin^{-1}(\epsilon/2\tau))^k \text{vol}(B_\epsilon)}$$

Let  $U = \cup_{x \in \bar{x}} B_\epsilon(x)$

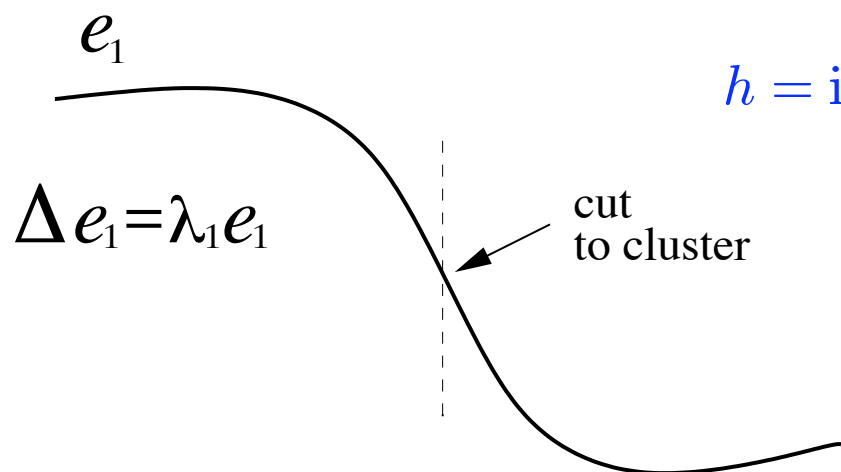
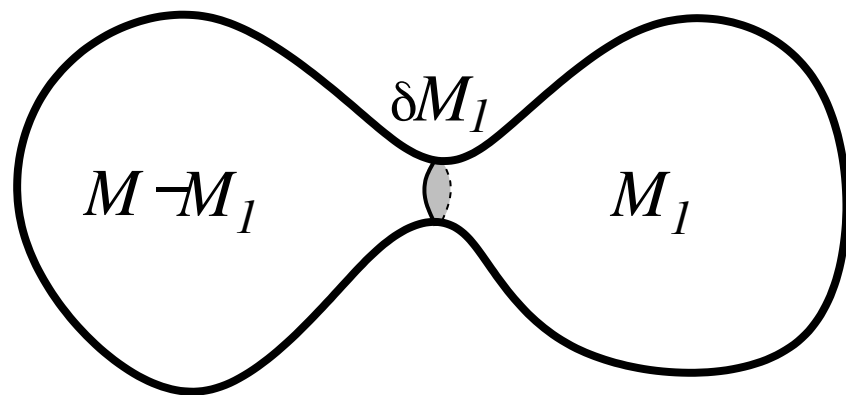
$$n > \beta(\log(\beta) + \log(\frac{1}{\delta}))$$

with prob.  $> 1 - \delta$ ,  
homology of  $U$  equals the homology of  $\mathcal{M}$

(Niyogi, Smale, Weinberger, 2004)

# Spectral Clustering

Isoperimetric inequalities. Cheeger constant.

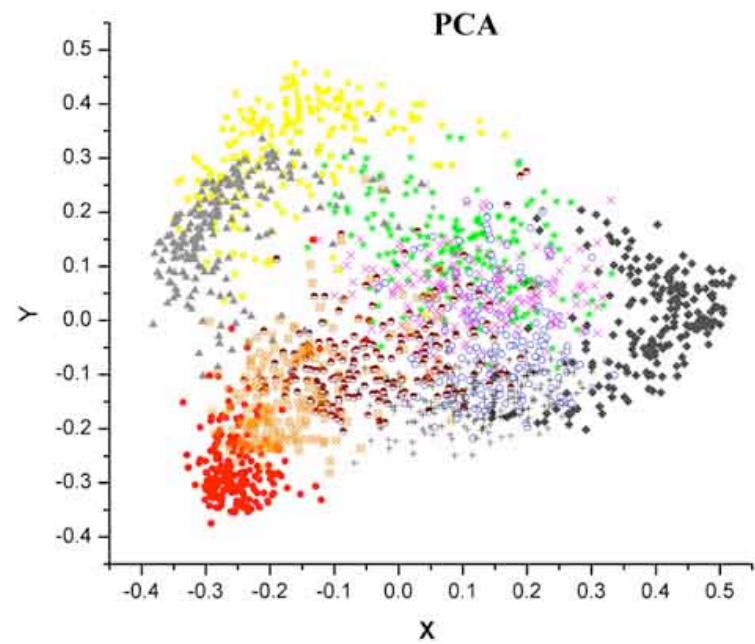
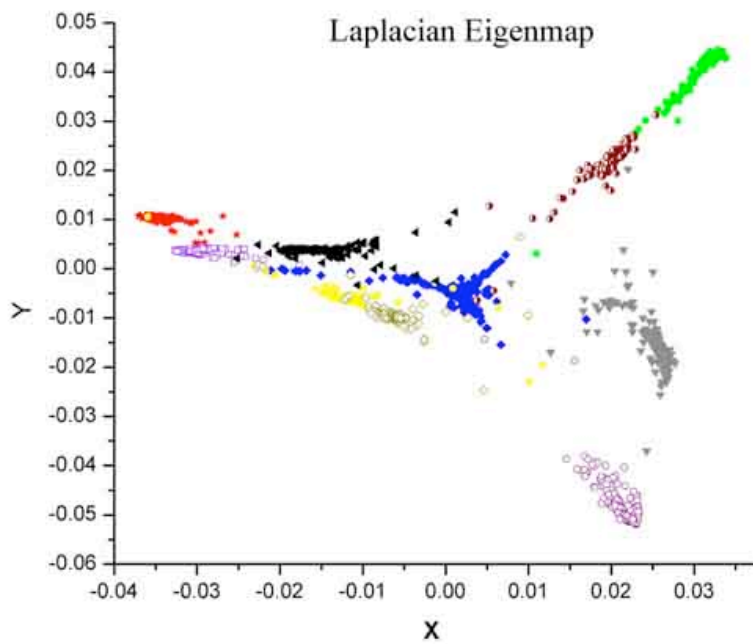


$$h = \inf \frac{\text{vol}^{n-1}(\delta \mathcal{M}_1)}{\min(\text{vol}^n(\mathcal{M}_1), \text{vol}^n(\mathcal{M} - \mathcal{M}_1))}$$

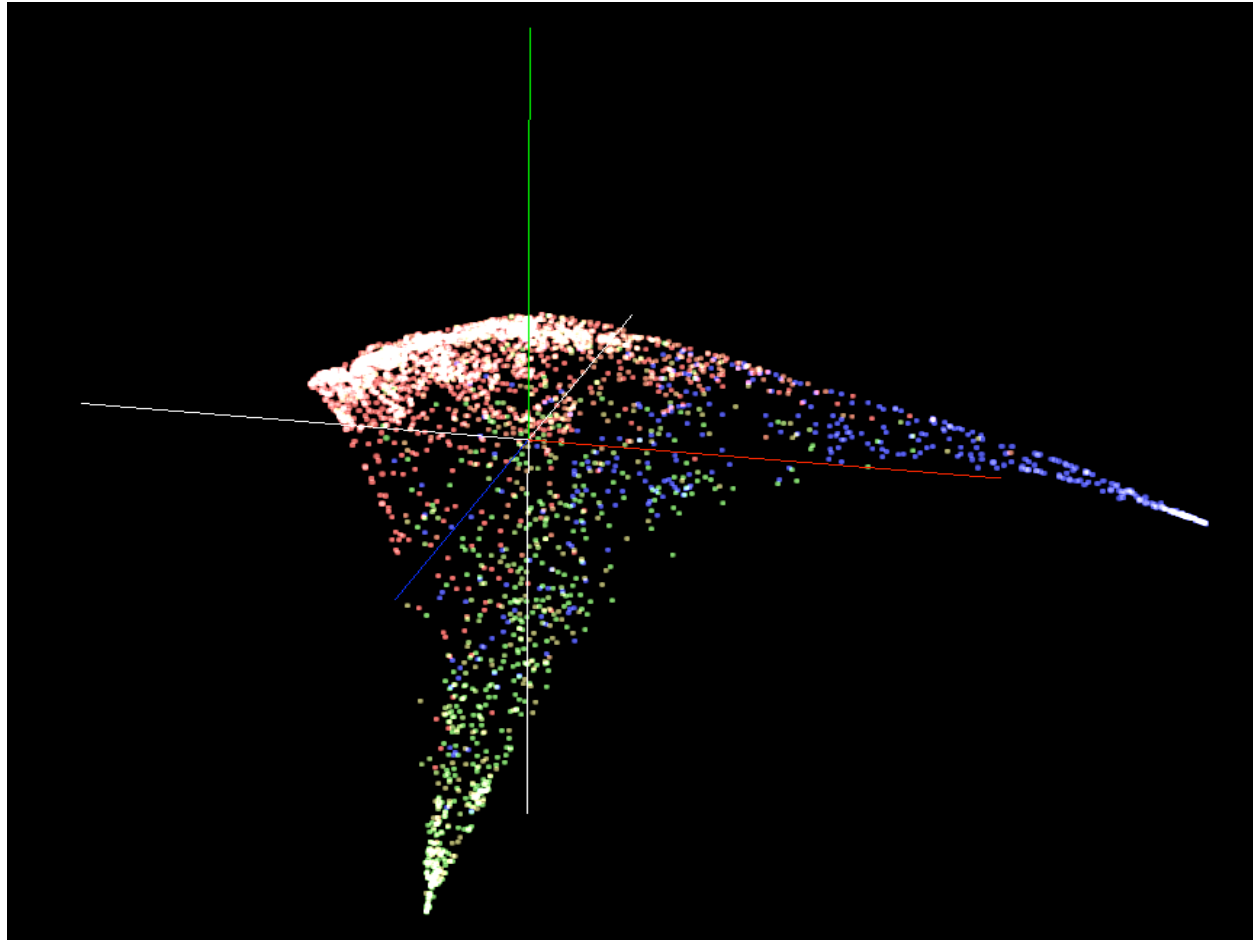
$$h \leq \frac{\sqrt{\lambda_1}}{2}$$

[Cheeger]

# Clustering Digits



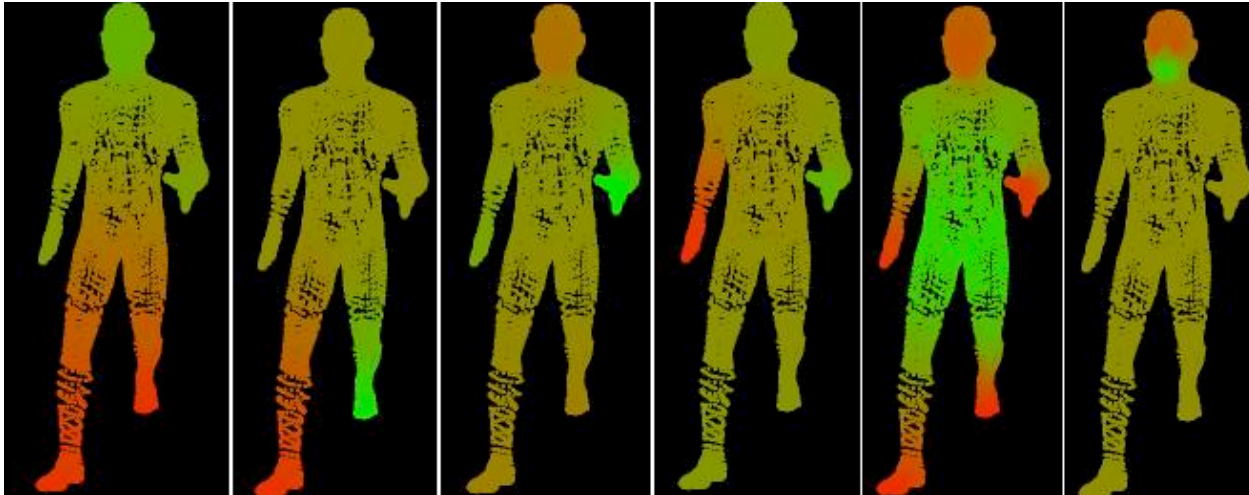
# Clustering Speech



# Computer Vision: Laplacian Eigenmaps

Machine vision: inferring joint angles.

Corazza, Andriacchi, Stanford Biomotion Lab, 05, Partiview, Surendran



Isometrically invariant representation.

# Important Issues

- How to handle noise theoretically and practically?
- How to choose the graph neighborhood correctly?
- How often do manifolds arise in natural data? What is the right metric on these manifolds?
- What are other ways in which one might utilize the geometry of natural distributions?
- Identify real problems where this approach can make a difference.
- Complexity estimates and provably correct algorithms rather than heuristics.