Dimension Reduction

James Sharpnack

Application to DARPA Intrusion Detection

SIPx	SPort	DIPx	DPort	Prot	PLen

Where

- SIP*x* = Source IP address nibble, where *x* = [1-4]. Four nibbles constitute the full source IP address
- SPort = Source Port number
- DIP*x* = Destination IP address nibble, where *x* = [1-4]. Four nibbles constitute the full destination IP address
- DPort = Destination Port number
- Prot = Protocol type: TCP, UDP or ICMP
- PLen = Packet length in bytes

[Labib, Khaled, and V. Rao Vemuri. "An Application of Principal Component Analysis to the Detection and Visualization of Computer Network Attacks.", 2007]

Application to DARPA Intrusion Detection

Table I : Description of DoS and Probe Attacks (Description des attaques de type DoS et Probe considérées)

Attack Name	Attack Description
Smurf (DoS)	Denial of Service ICMP echo reply flood
Neptune (DoS)	SYN flood Denial of Service on one or more ports
IPsweep (Probe)	Surveillance sweep performing either a port sweep or ping on multiple host addresses
Portsweep (Probe)	Surveillance sweep through many ports to determine which services are supported on a single host

[Labib, Khaled, and V. Rao Vemuri. "An Application of Principal Component Analysis to the Detection and Visualization of Computer Network Attacks.", 2007]

Principal Component Analysis

Setting, spectral decomposition, PCA

Application to DARPA Intrusion Detection

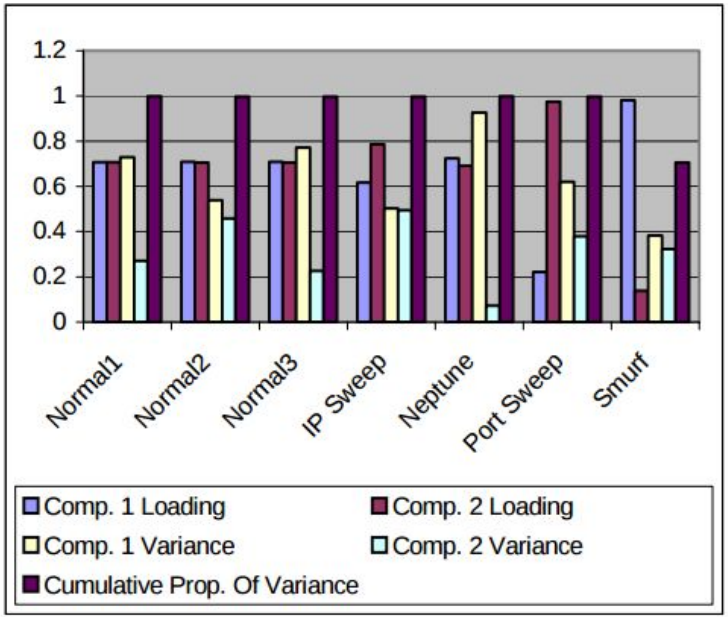


Figure 1: Component Loading and Variance (Charge des composants et variance)

[Labib, Khaled, and V. Rao Vemuri. "An Application of Principal Component Analysis to the Detection and Visualization of Computer Network Attacks.", 2007]

Application to image compression

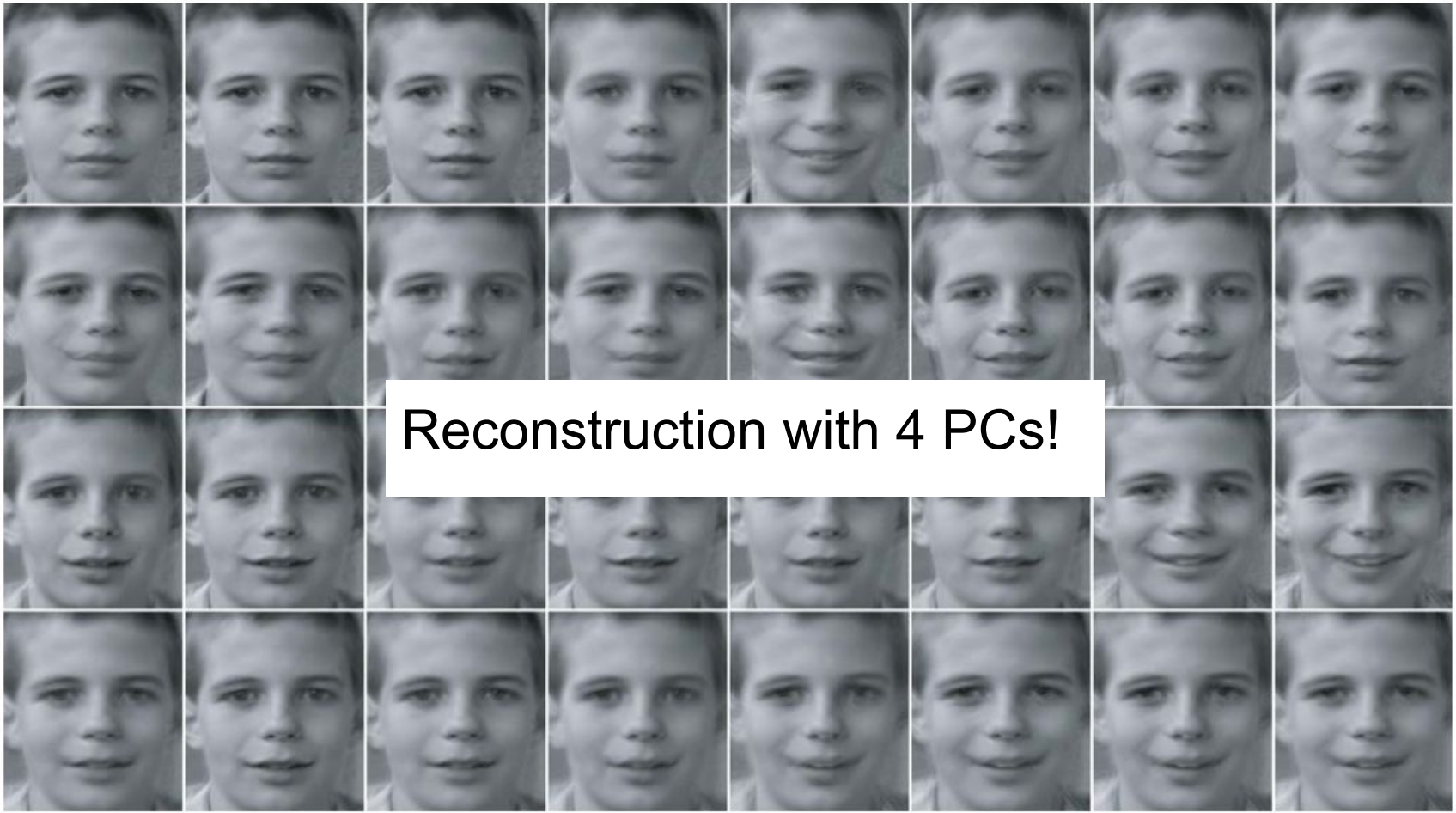
Image segmentation by color: n x n x L bits to store (L=log(k), k is # of clusters)

Can we write image as an n x n matrix then approximate it with rank k matrix?

n x k x # bits

[Sonka at al. “Image Processing, Analysis, and Machine Vision”, 2014]

[Sonka at al. “Image Processing, Analysis, and Machine Vision”, 2014]

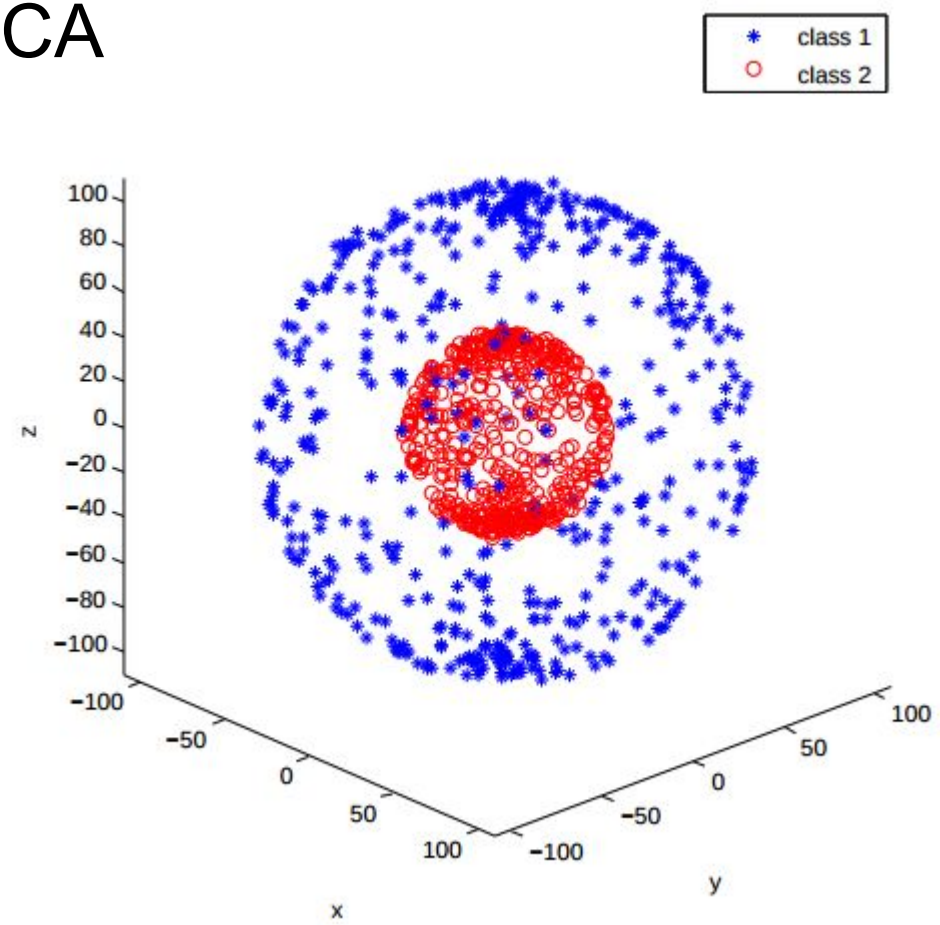


[Sonka at al. “Image Processing, Analysis, and Machine Vision”, 2014]

Figure 6. The head has been located—the image in the box is sent to the face recognition process. Also shown is the path of the head tracked over several previous frames.

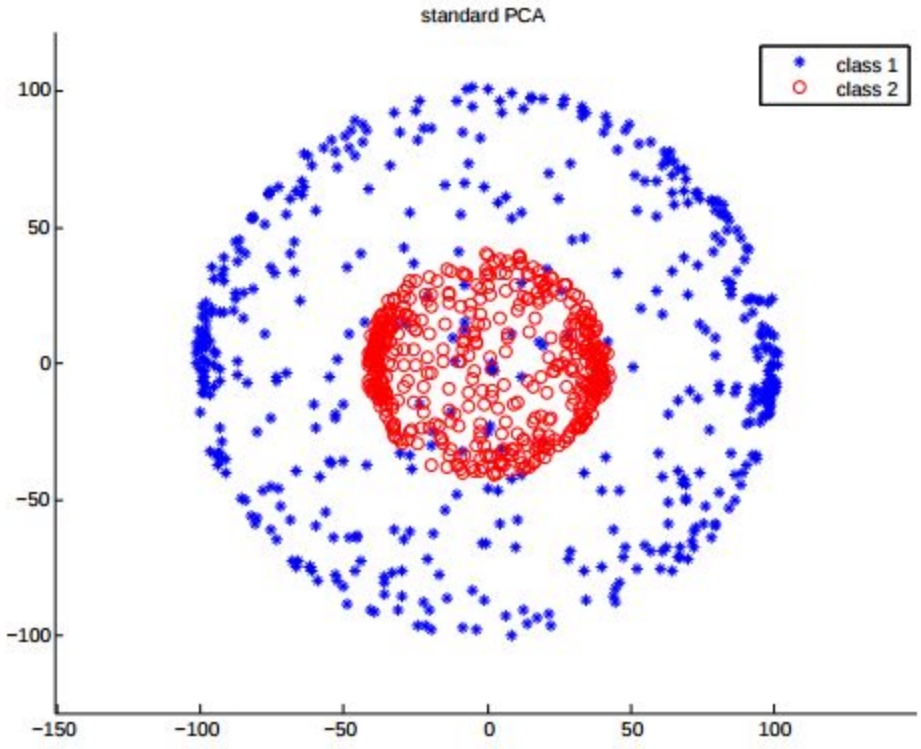
[Turk, Pentland. “Eigenfaces for Recognition”, 1991]

Non-linear PCA



[Wang, 2014]

Non-linear PCA

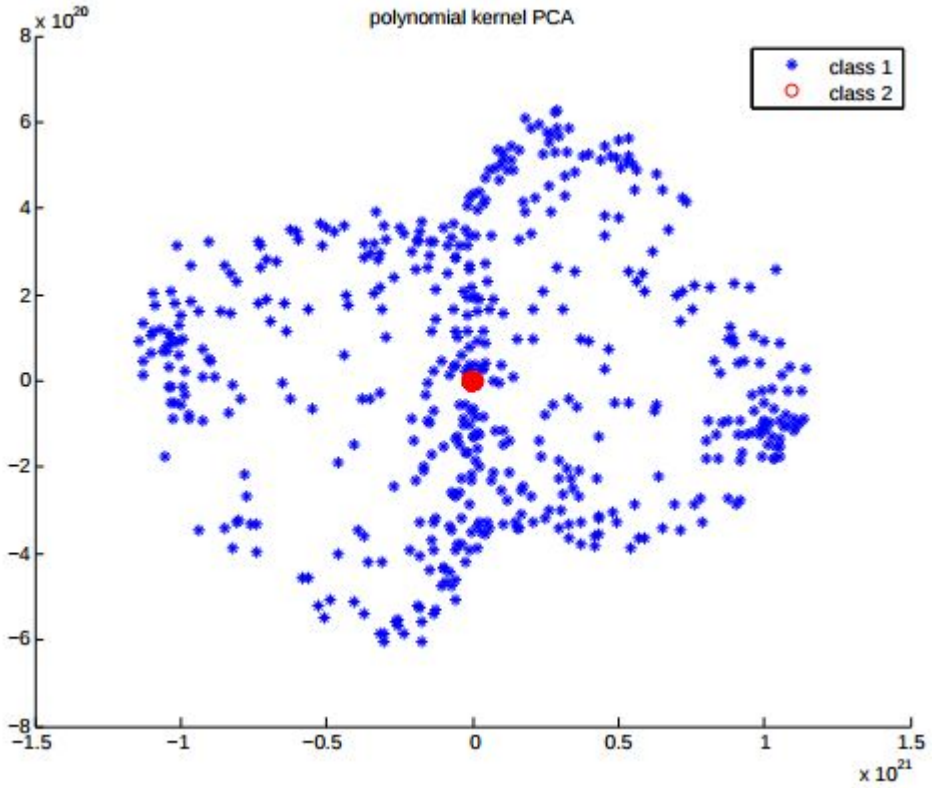


[Wang, 2014]

Kernel PCA

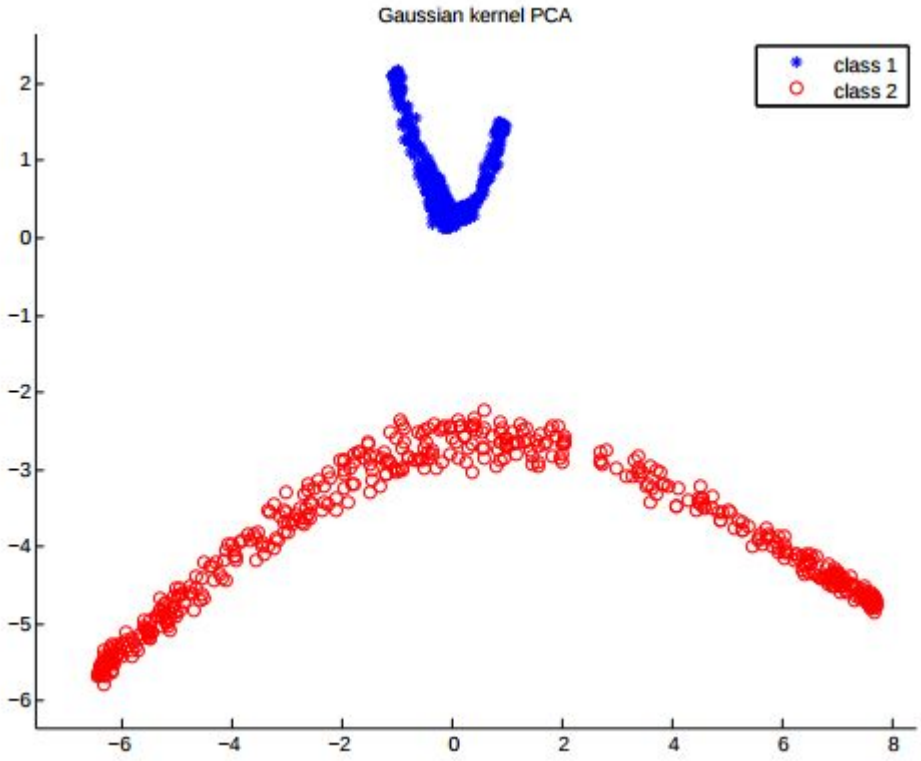
Method

Poly-kernel PCA



[Wang, 2014]

RBF-kernel PCA



[Wang, 2014]

USPS Dataset

Vectorized images (256 dim)

Linear v RBF kernel PCA

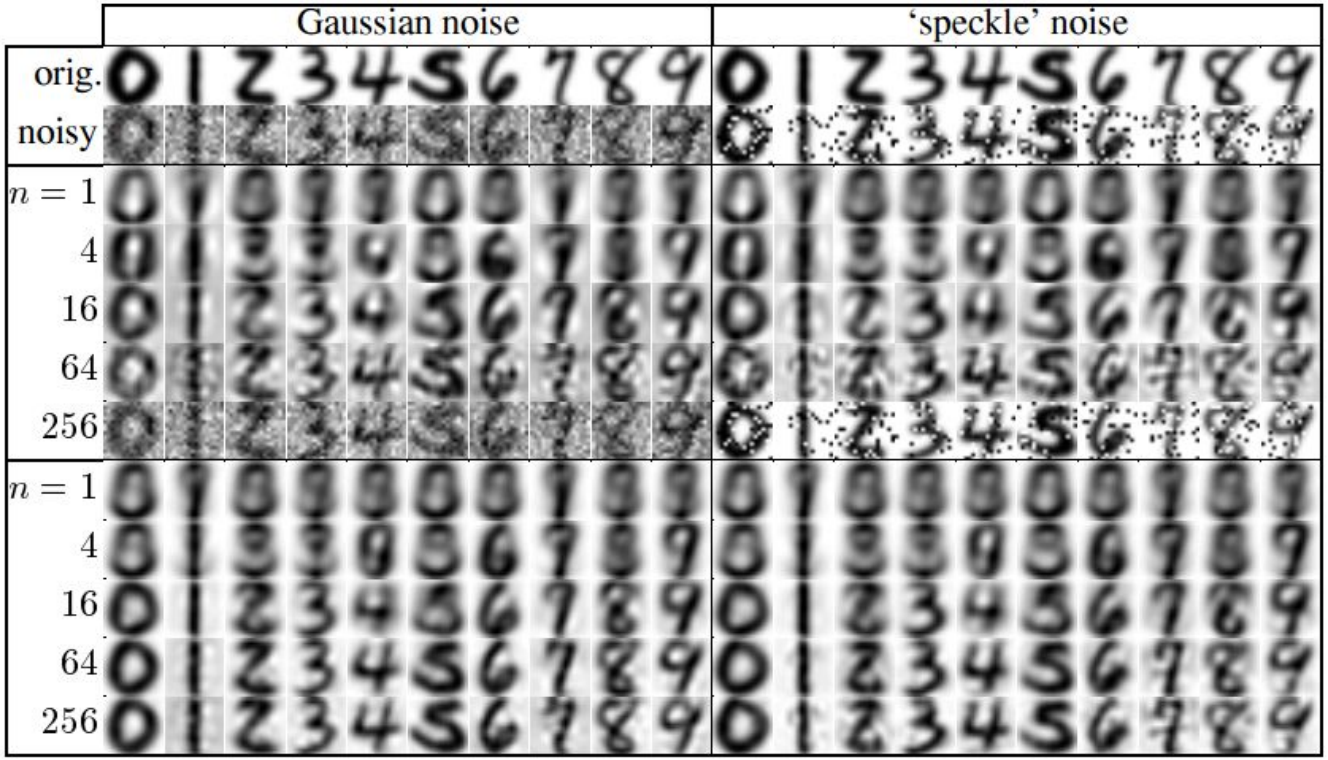


Figure 4: De-Noising of USPS data (see text). The left half shows: *top:* the first occurrence of each digit in the test set, *second row:* the upper digit with additive Gaussian noise ($\sigma = 0.5$), *following five rows:* the reconstruction for linear PCA using $n = 1, 4, 16, 64, 256$ components, and, *last five rows:* the results of our approach using the same number of components. In the right half we show the same but for 'speckle' noise with probability $p = 0.4$.

[Mika et al., 1999]

Laplacian Eigenmaps

Graph, Laplacian

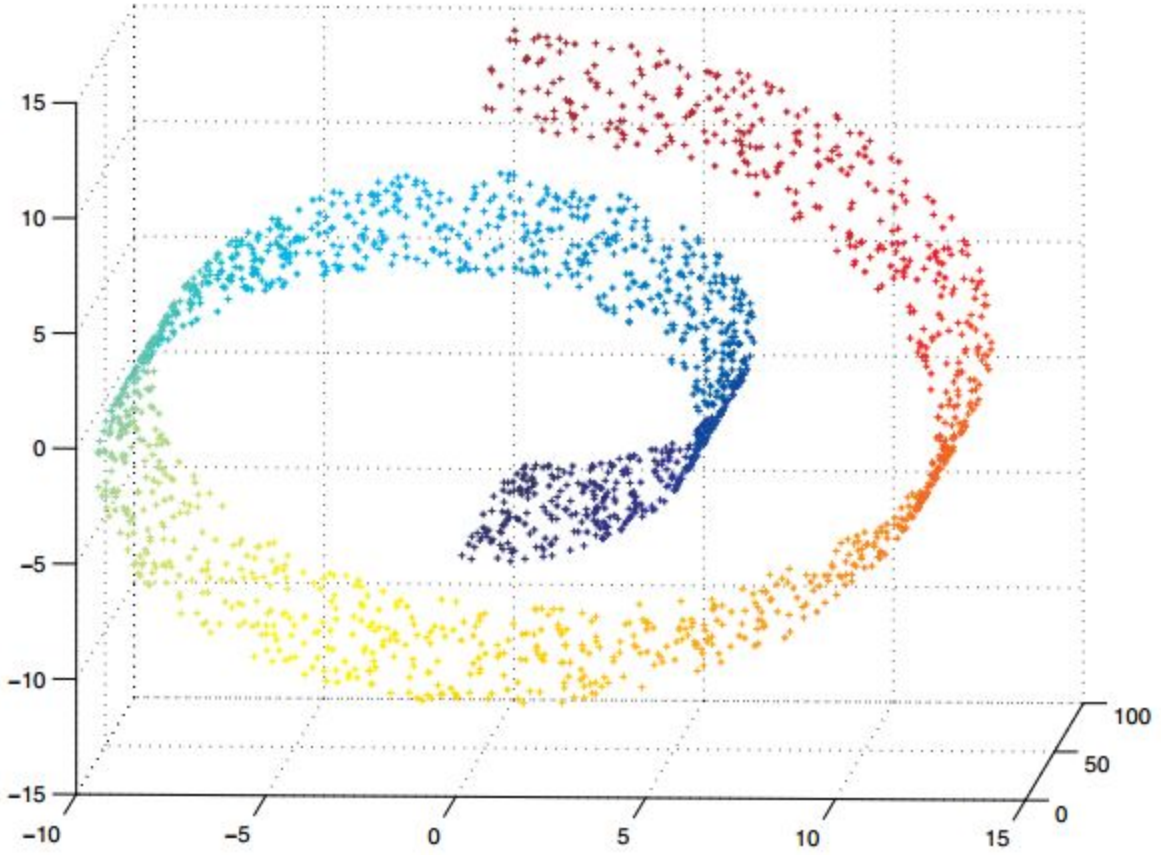


Figure 1: 2000 Random data points on the swiss roll.

[Belkin, Niyogi,

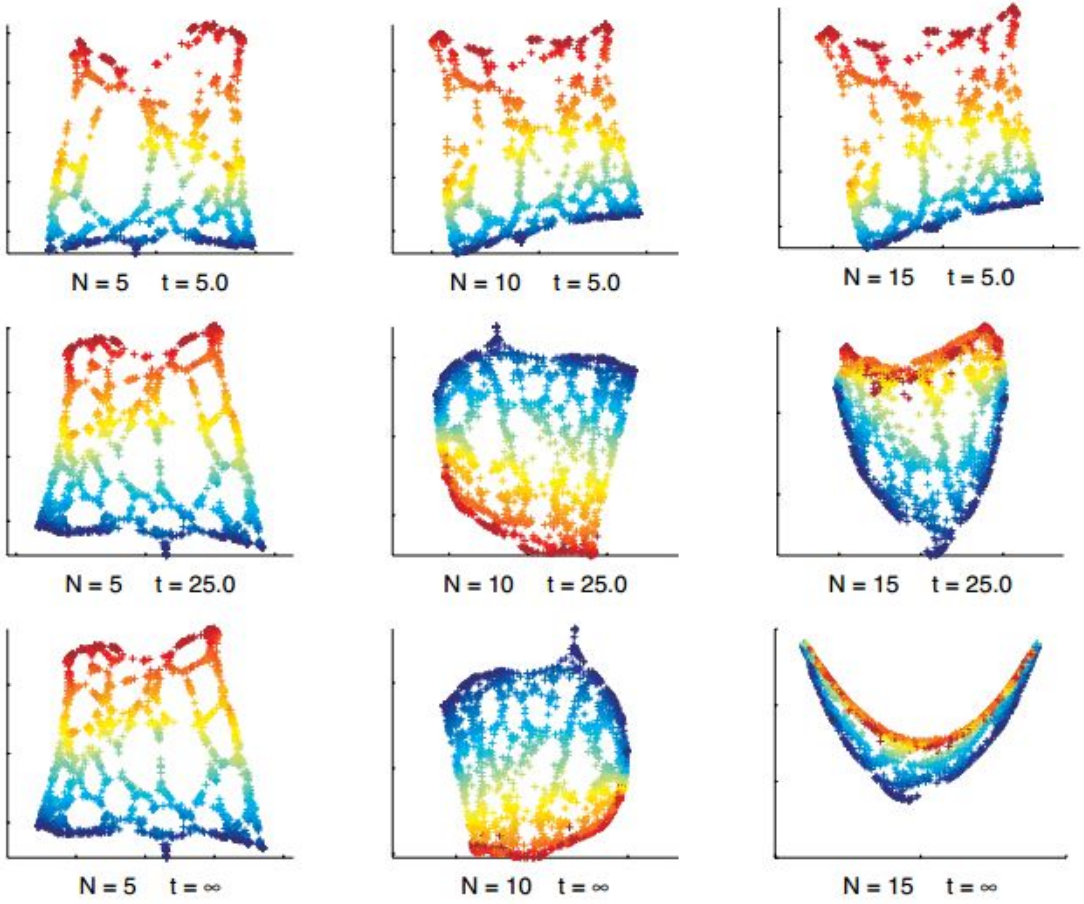


Figure 2: Two-dimensional representations of the swiss roll data, for different values of the number of nearest neighbors N and the heat kernel parameter t. $t = \infty$ corresponds to the discrete weights.

Brown corpus

300 most frequent words in a body of documents

feature vector: how frequent is word j before word i, and vice versa (600 dim)

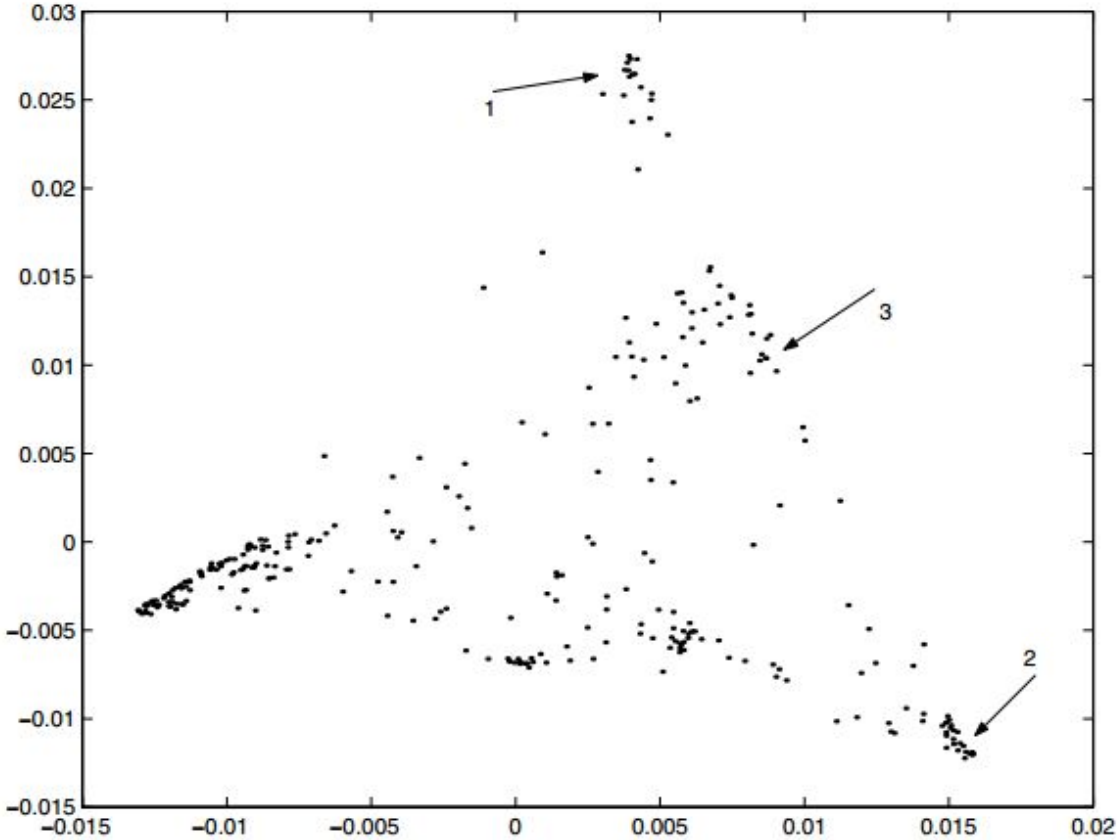


Figure 4: The 300 most frequent words of the Brown corpus represented in the spectral domain.

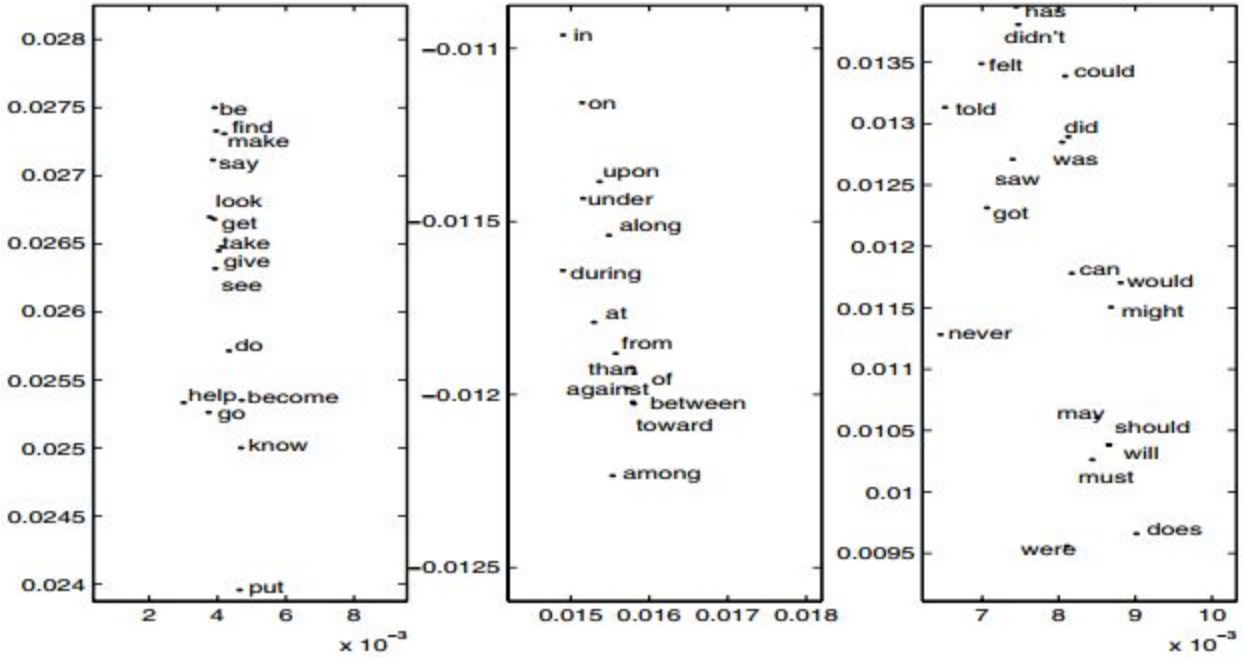


Figure 5: Fragments labeled by arrows: (left) infinitives of verbs, (middle) prepositions, and (right) mostly modal and auxiliary verbs. We see that syntactic structure is well preserved.

Semisupervised learning

Suppose that not all points are labelled (x_i, y_i) and some are not (x_j).

We can use the learned manifold to denoise and label the unlabelled data...

Protein function prediction

Table 1: 13 CYGD functional classes

	Classes
1	Metabolism
2	Energy
3	Cell cycle and DNA processing
4	Transcription
5	Protein synthesis
6	Protein fate
7	Cellular transportation and transportation mechanism
8	Cell rescue, defense and virulence
9	Interaction with cell environment
10	Cell fate
11	Control of cell organization
12	Transport facilitation
13	Others

[Tran, 2013]

Protein function prediction

Functional Classes	Accuracy Performance Measures (%) Network $W^{(1)}$		
	Normalized	Random Walk	Un-normalized
1	64.24	63.96	**64.30**
2	71.01	71.07	**71.13**
3	63.88	63.66	**63.91**
4	**65.55**	65.41	65.47
5	71.35	**71.46**	71.24
6	66.95	66.69	**67.11**
7	**67.89**	67.70	67.84
8	69.29	69.29	**69.31**
9	71.49	71.40	**71.52**
10	65.30	65.47	**65.50**

[Tran, 2013]