UNIVERSITY OF MINNESOTA TWIN CITIES

Linear algebra methods for data mining with applications to materials

Yousef Saad

Department of Computer Science and Engineering

University of Minnesota

SIAM 2012 Annual Meeting

Twin Cities, July 13th, 2012

Introduction: a few factoids

- Data is growing exponentially at an "alarming" rate:
  - 90% of data in world today was created in last two years
  - Every day, 2.3 Million terabytes (2.3×10^{18} bytes) created
- "Big data" term coined to reflect this trend
- Mixed blessing: Opportunities & big challenges.
- Trend is re-shaping & energizing many research areas ...
- ... including my own: numerical linear algebra

Introduction: What is data mining?

Set of methods and tools to extract meaningful information or patterns from data. Broad area : data analysis, machine learning, pattern recognition, information retrieval, ...

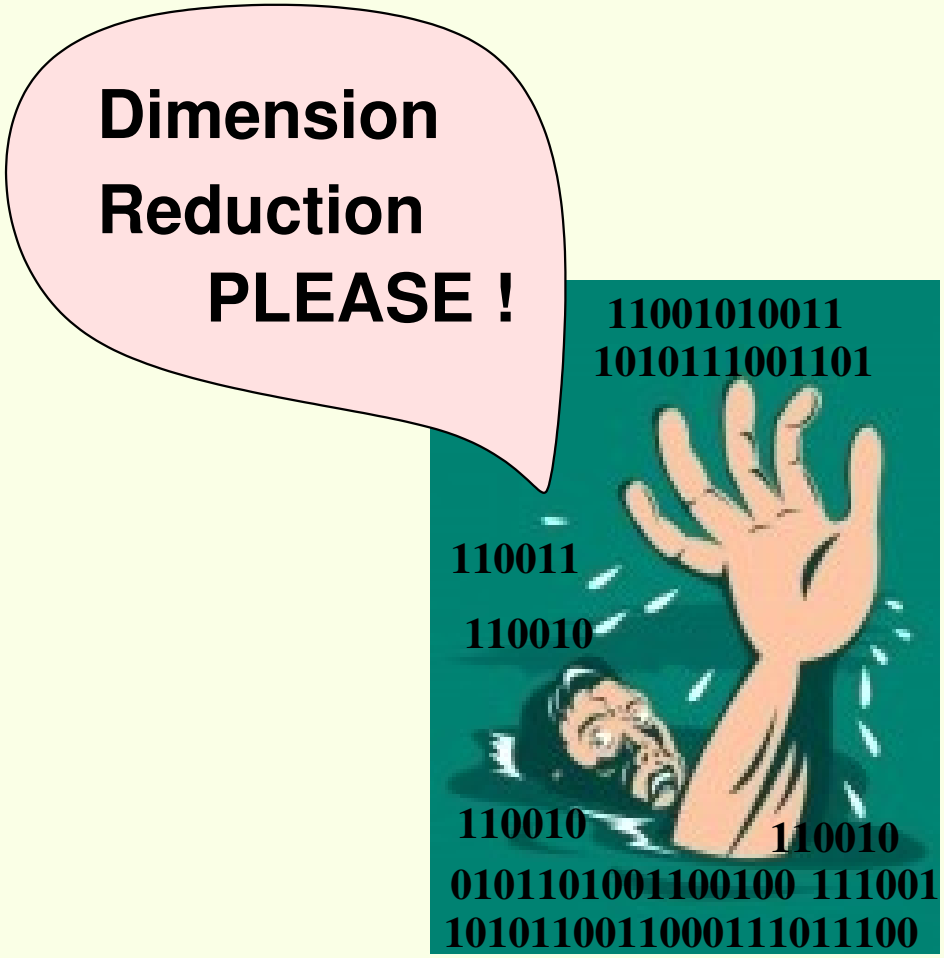
- Tools used: linear algebra; Statistics; Graph theory; Approximation theory; Optimization; ...

- This talk: brief introduction – with emphasis on linear algebra viewpoint

- + our initial work on materials.

- Big emphasis on "Dimension reduction methods"

Drowning in data



Picture modified from http://www.123rf.com/photo_7238007_man-drowning-reaching-out-for-help.html

Major tool of Data Mining: Dimension reduction

➤ Goal is not as much to reduce size (& cost) but to:

- Reduce noise and redundancy in data before performing a task [e.g., classification as in digit/face recognition]
- Discover 'important features' or 'patterns'

➤ Map data to: Preserve proximity? Maximize variance? Preserve a certain graph?

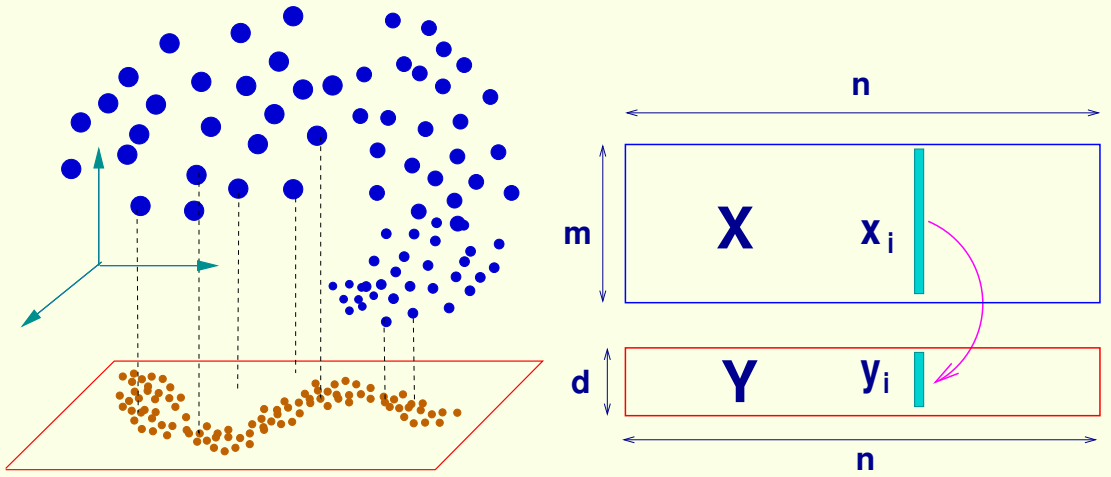
➤ Other term used: feature extraction [general term for low dimensional representations of data]

The problem of Dimension Reduction

➤ Given $d \ll m$ find a mapping Φ:

$$\Phi : x \in \mathbb{R}^m \longrightarrow y \in \mathbb{R}^d$$

Practically: Given: $X = [x_1, \cdots, x_n] \in \mathbb{R}^{m \times n}$, find a low-dimens. representation $Y = [y_1, \cdots, y_n] \in \mathbb{R}^{d \times n}$ of X



Φ may be linear, or nonlinear (implicit). Linear case ($W \in \mathbb{R}^{m \times d}$):

$$Y = W^\top X$$

Example: Principal Component Analysis (PCA)

In *Principal Component Analysis* $\boldsymbol{W}$ is computed to maximize variance of projected data:

$$\max_{\substack{\boldsymbol{W} \in \mathbb{R}^{m \times d} \\ \boldsymbol{W}^\top \boldsymbol{W} = \boldsymbol{I}}} \sum_{i=1}^{d} \left\| \boldsymbol{y}_i - \frac{1}{n} \sum_{j=1}^{n} \boldsymbol{y}_j \right\|_2^2, \quad \boldsymbol{y}_i = \boldsymbol{W}^\top \boldsymbol{x}_i.$$

- Leads to maximizing

$$\text{Tr}\left[\boldsymbol{W}^\top (\boldsymbol{X} - \boldsymbol{\mu}\boldsymbol{e}^\top)(\boldsymbol{X} - \boldsymbol{\mu}\boldsymbol{e}^\top)^\top \boldsymbol{W}\right], \quad \boldsymbol{\mu} = \tfrac{1}{n}\Sigma_{i=1}^{n} \boldsymbol{x}_i$$

- Solution $\boldsymbol{W} = \{$ dominant eigenvectors $\}$ of the covariance matrix == Set of left singular vectors of $\bar{\boldsymbol{X}} = \boldsymbol{X} - \boldsymbol{\mu e}^\top$

SVD:

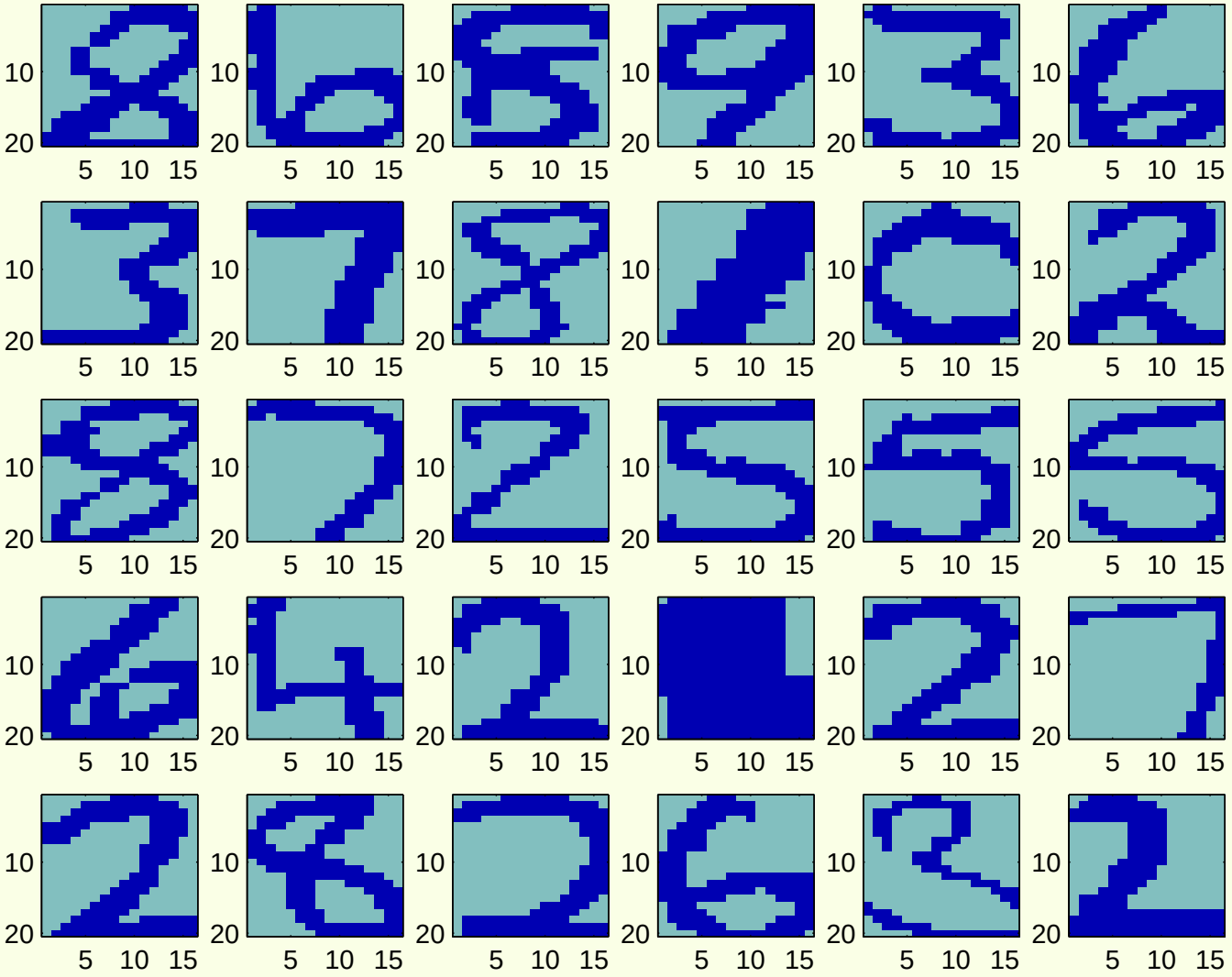
$$\bar{\boldsymbol{X}} = \boldsymbol{U\Sigma V}^\top, \quad \boldsymbol{U}^\top\boldsymbol{U} = \boldsymbol{I}, \quad \boldsymbol{V}^\top\boldsymbol{V} = \boldsymbol{I}, \quad \boldsymbol{\Sigma} = \text{Diag}$$

- Optimal $\boldsymbol{W} = \boldsymbol{U}_d \equiv$ matrix of first $\boldsymbol{d}$ columns of $\boldsymbol{U}$
- Solution $\boldsymbol{W}$ also minimizes 'reconstruction error' ..

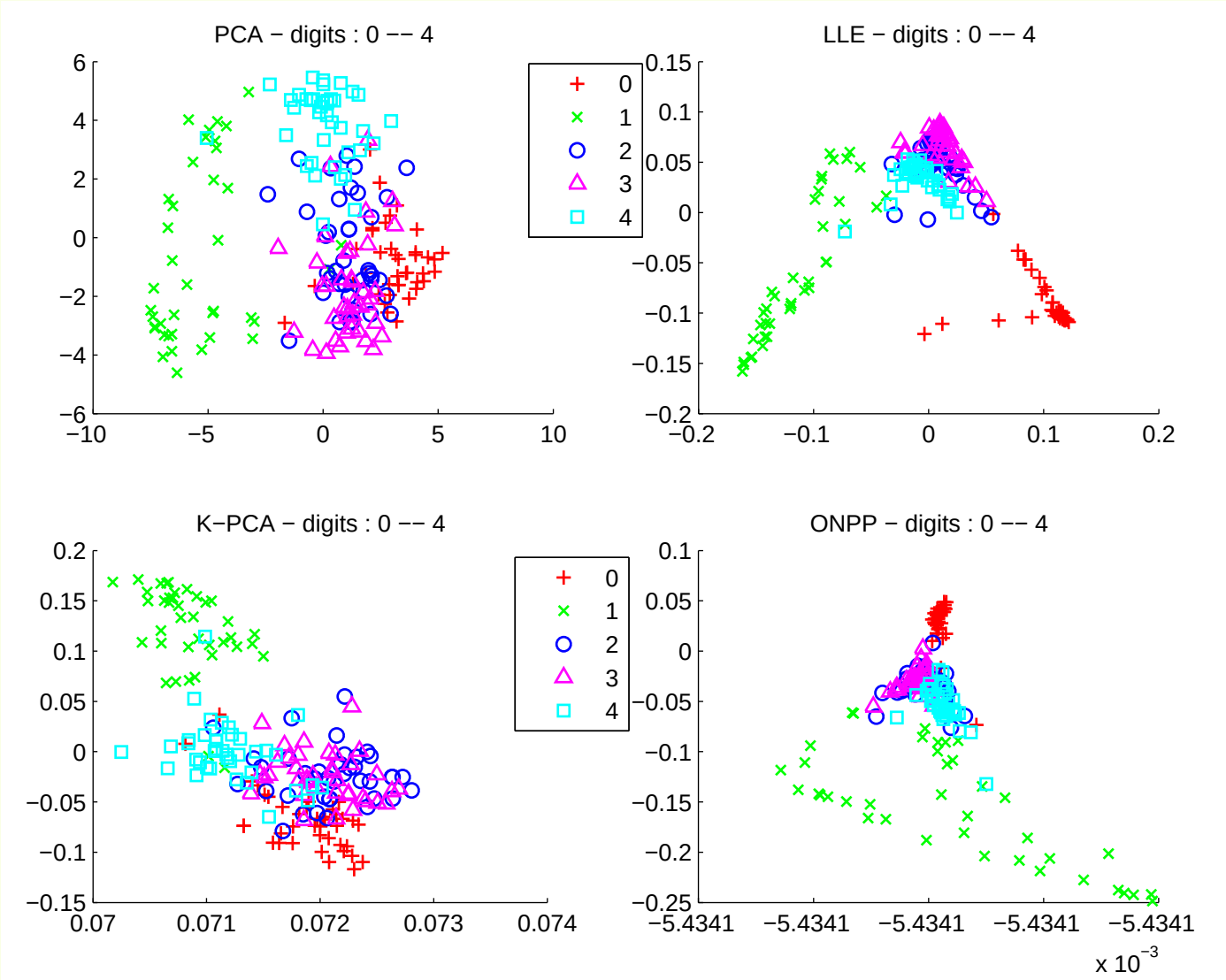
$$\sum_i \|\boldsymbol{x}_i - \boldsymbol{W}\boldsymbol{W}^T\boldsymbol{x}_i\|^2 = \sum_i \|\boldsymbol{x}_i - \boldsymbol{W}\boldsymbol{y}_i\|^2$$

- In some methods recentering to zero is not done, i.e., $\bar{\boldsymbol{X}}$ replaced by $\boldsymbol{X}$.

Example : Digit images (a random sample of 30)



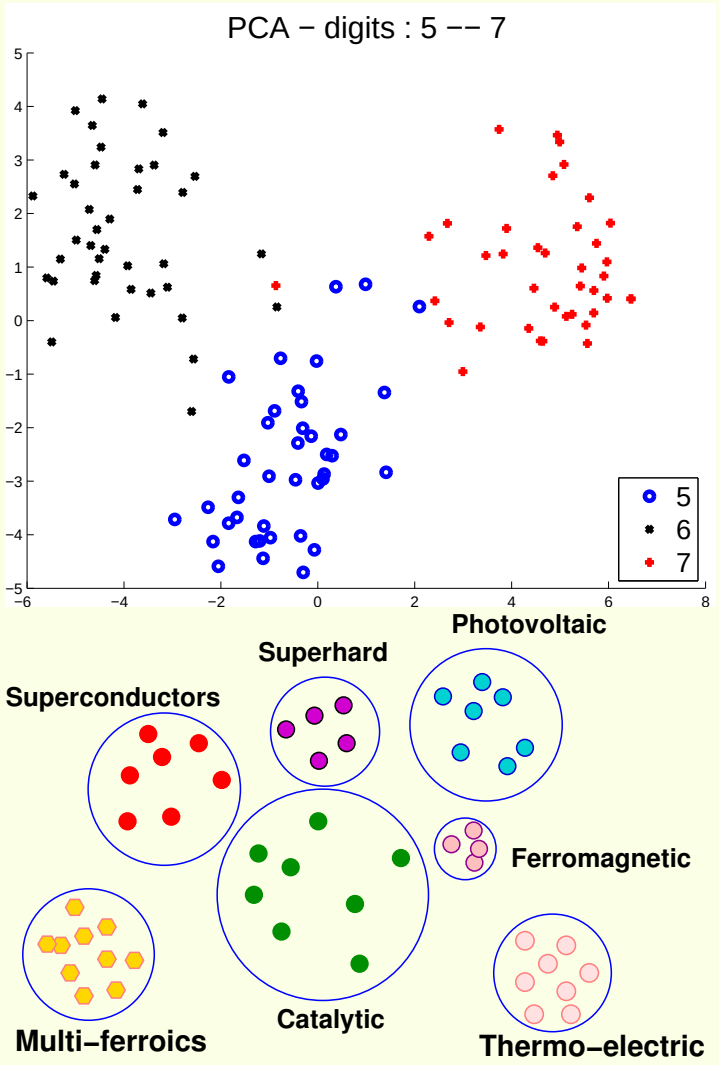
2-D 'reductions':



Unsupervised learning

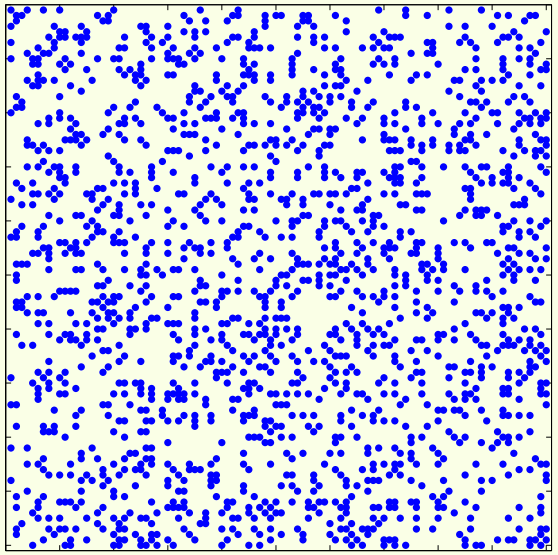
"Unsupervised learning" : methods that do not exploit known labels

- Example of digits: perform a 2-D projection
- Images of same digit tend to cluster (more or less)
- Such 2-D representations are popular for visualization
- Can also try to find natural clusters in data, e.g., in materials
- Basic clusterning technique: K-means



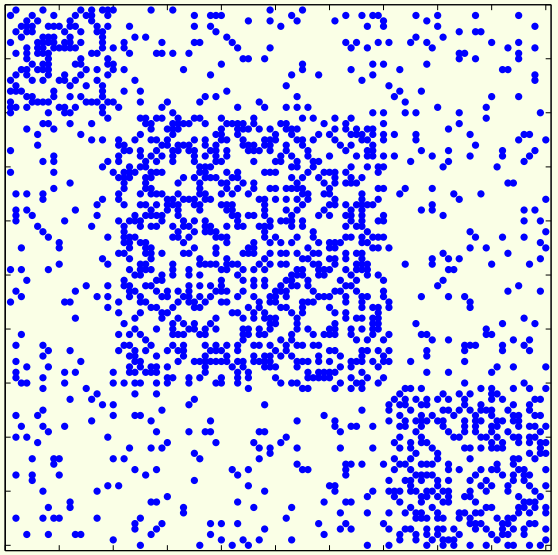
Example: Sparse Matrices viewpoint (J. Chen & YS '09)

- Communities modeled by an 'affinity' graph [e.g., 'user A sends frequent e-mails to user B']
- Adjacency Graph represented by a sparse matrix



← Original matrix

Goal: Find ordering so blocks are as dense as possible →



- Use 'blocking' techniques for sparse matrices
- Advantage of this viewpoint: need not know # of clusters.

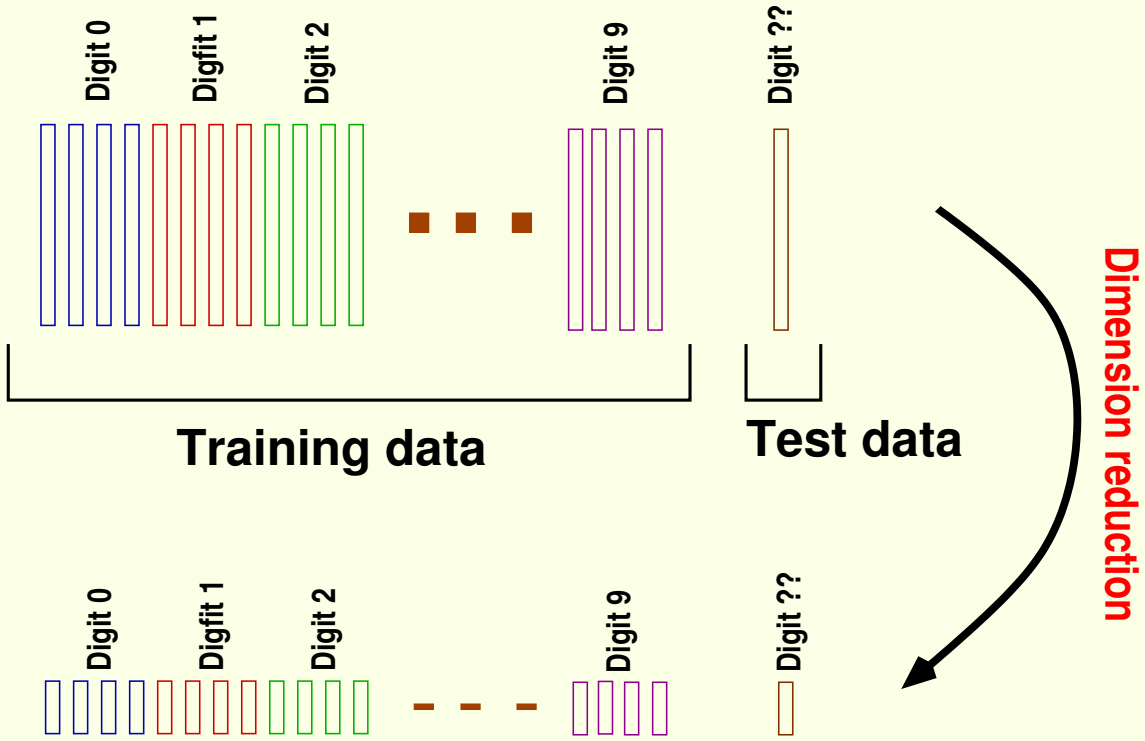
[data: `www-personal.umich.edu/~mejn/netdata/`]

Supervised learning: classification

- Best illustration: written digits recognition example

Given: a set of labeled samples (training set), and an (unlabeled) test image.

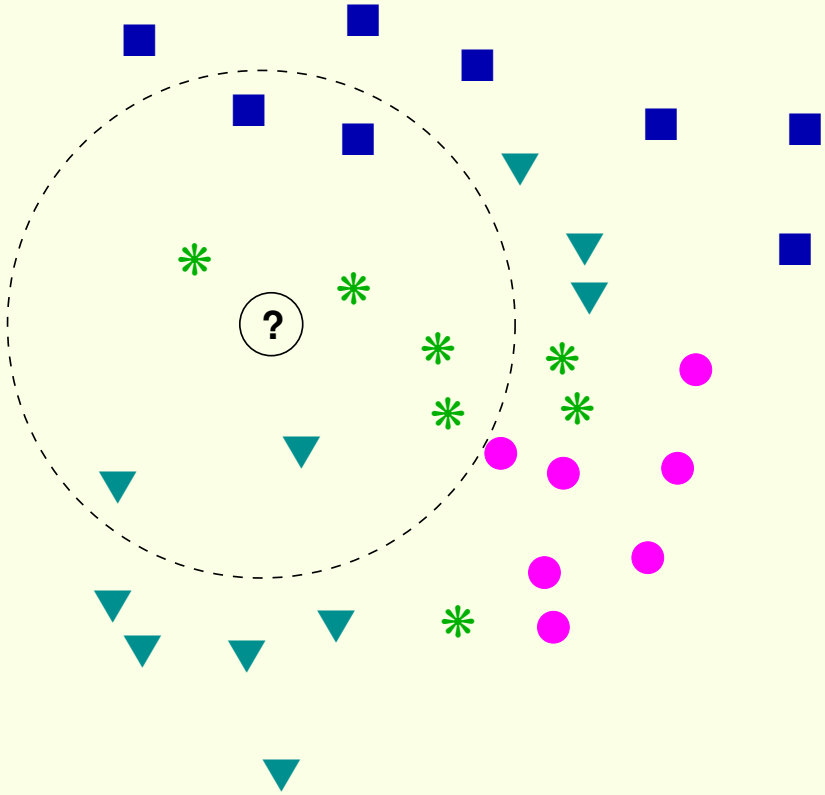
Problem: find label of test image



- Roughly speaking: we seek dimension reduction so that recognition is 'more effective' in low-dim. space

K-nearest neighbors (KNN) classification

- Idea of a voting system: get distances between test sample and training samples
- Get the k nearest neighbors (here $k = 8$)
- Predominant class among these k items is assigned to the test sample ("$*$" here)



Fisher's Linear Discriminant Analysis (LDA)

Goal: Use label information to define a good projector, i.e., one that can 'discriminate' well between given classes

➤ Define "between scatter": a measure of how well separated two distinct classes are.

➤ Define "within scatter": a measure of how well clustered items of the same class are.

➤ Objective: make "between scatter" measure large and "within scatter" small.

Idea: Find projector that maximizes the ratio of the "between scatter" measure over "within scatter" measure

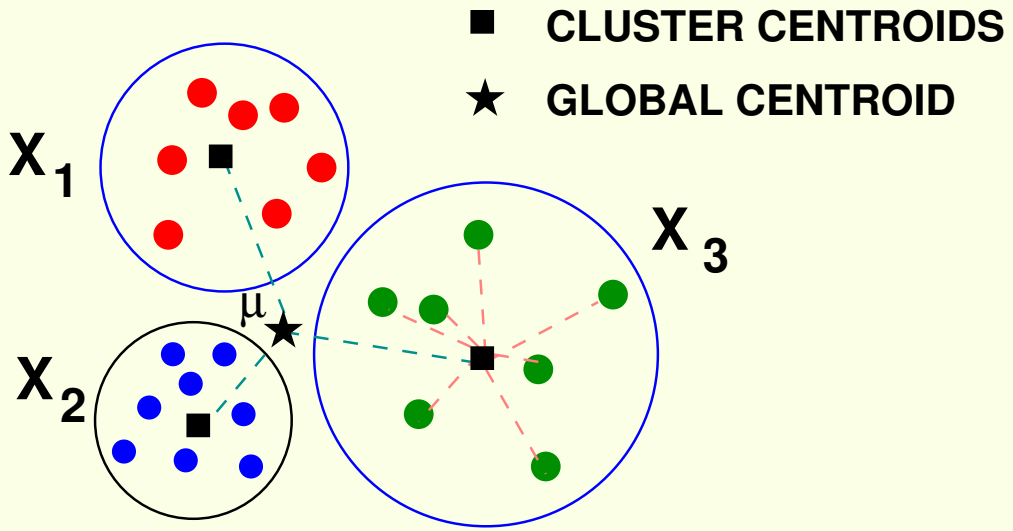
Define:

$$S_B = \sum_{k=1}^{c} n_k (\mu^{(k)} - \mu)(\mu^{(k)} - \mu)^T,$$

$$S_W = \sum_{k=1}^{c} \sum_{x_i \in X_k} (x_i - \mu^{(k)})(x_i - \mu^{(k)})^T$$

Where:

- μ = mean (X)
- $\mu^{(k)}$ = mean (X_k)
- $X_k = k$-th class
- $n_k = |X_k|$



➤ Consider 2nd moments for a vector a:

$$a^T S_B a = \sum_{i=1}^{c} n_k \, |a^T(\mu^{(k)} - \mu)|^2,$$

$$a^T S_W a = \sum_{k=1}^{c} \sum_{x_i \in X_k} |a^T(x_i - \mu^{(k)})|^2$$

➤ $a^T S_B a \equiv$ weighted variance of projected μ_j's

➤ $a^T S_W a \equiv$ w. sum of variances of projected classes X_j's

➤ LDA projects the data so as to maximize the ratio of these two numbers:

$$\max_a \frac{a^T S_B a}{a^T S_W a}$$

➤ Optimal a = eigenvector associated with the largest eigenvalue of: $S_B u_i = \lambda_i S_W u_i$.

LDA – Extension to arbitrary dimensions

- Criterion: maximize the ratio of two traces:

$$\frac{\text{Tr}\,[U^T S_B U]}{\text{Tr}\,[U^T S_W U]}$$

- Constraint: $U^T U = I$ (orthogonal projector).
- Reduced dimension data: $Y = U^T X$.

Common viewpoint: hard to maximize, therefore ...

- ... alternative: Solve instead the ('easier') problem:

$$\max_{U^T S_W U = I} \text{Tr}\,[U^T S_B U]$$

- Solution: largest eigenvectors of $S_B u_i = \lambda_i S_W u_i$.

LDA – Extension to arbitrary dimensions (cont.)

➤ Consider the original problem:

$$\max_{U \in \mathbb{R}^{n\times p},\, U^T U = I} \frac{\mathrm{Tr}\,[U^T A U]}{\mathrm{Tr}\,[U^T B U]}$$

Let A, B be symmetric & assume that B is semi-positive definite with $\mathrm{rank}(B) > n - p$. Then $\mathrm{Tr}\,[U^T A U]/\mathrm{Tr}\,[U^T B U]$ has a finite maximum value ρ_*. The maximum is reached for a certain U_* that is unique up to unitary transforms of columns.

➤ Consider the function:

$$f(\rho) = \max_{V^T V = I} \mathrm{Tr}\,[V^T (A - \rho B) V]$$

➤ Call $V(\rho)$ the maximizer.

➤ Note: $V(\rho)$ = Set of eigenvectors - not unique

- Define $G(\rho) \equiv A - \rho B$ and its n eigenvalues:

$$\mu_1(\rho) \geq \mu_2(\rho) \geq \cdots \geq \mu_n(\rho) \, .$$

- Clearly:

$$f(\rho) = \mu_1(\rho) + \mu_2(\rho) + \cdots + \mu_p(\rho) \, .$$

- Can express this differently. Define eigenprojector:

$$P(\rho) = V(\rho) V(\rho)^T$$

- Then:

$$\begin{aligned} f(\rho) &= \text{Tr}\left[V(\rho)^T G(\rho) V(\rho)\right] \\ &= \text{Tr}\left[G(\rho) V(\rho) V(\rho)^T\right] \\ &= \text{Tr}\left[G(\rho) P(\rho)\right]. \end{aligned}$$

➤ Recall [e.g. Kato '65] that:

$$P(\rho) = \frac{-1}{2\pi i} \int_{\Gamma} (G(\rho) - zI)^{-1}\, dz$$

Γ is a smooth curve containing the p eivenvalues of interest

➤ Hence: $f(\rho) = \frac{-1}{2\pi i} \mathrm{Tr} \int_{\Gamma} G(\rho)(G(\rho) - zI)^{-1}\, dz = \ldots$

$$= \frac{-1}{2\pi i} \mathrm{Tr} \int_{\Gamma} z(G(\rho) - zI)^{-1}\, dz$$

➤ With this, can prove :

1. f is a non-increasing function of ρ;
2. $f(\rho) = 0$ iff $\rho = \rho_*$;
3. $f'(\rho) = -\mathrm{Tr}\,[V(\rho)^T B V(\rho)]$

Can now use Newton's method.

$$\rho_{new} = \rho - \frac{\text{Tr}\,[V(\rho)^T(A-\rho B)V(\rho)]}{-\text{Tr}\,[V(\rho)^T B V(\rho)]} = \frac{\text{Tr}\,[V(\rho)^T A V(\rho)]}{\text{Tr}\,[V(\rho)^T B V(\rho)]}$$

- Newton's method to find the zero of $f \equiv$ a fixed point iteration with $$g(\rho) = \frac{\text{Tr}\,[V^T(\rho) A V(\rho)]}{\text{Tr}\,[V^T(\rho) B V(\rho)]},$$
- Idea: Compute $V(\rho)$ by a Lanczos-type procedure
- Note: Standard problem - [not generalized] $\longrightarrow$ inexpensive!
- See T. Ngo, M. Bellalij, and Y.S. 2010 for details

- Recent papers advocated similar or related techniques

[1] C. Shen, H. Li, and M. J. Brooks, A convex programming approach to the trace quotient problem. In *ACCV (2)* – 2007.

[2] H. Wang, S.C. Yan, D.Xu, X.O. Tang, and T. Huang. Trace ratio vs. ratio trace for dimensionality reduction. In *IEEE Conference on Computer Vision and Pattern Recognition, 2007*

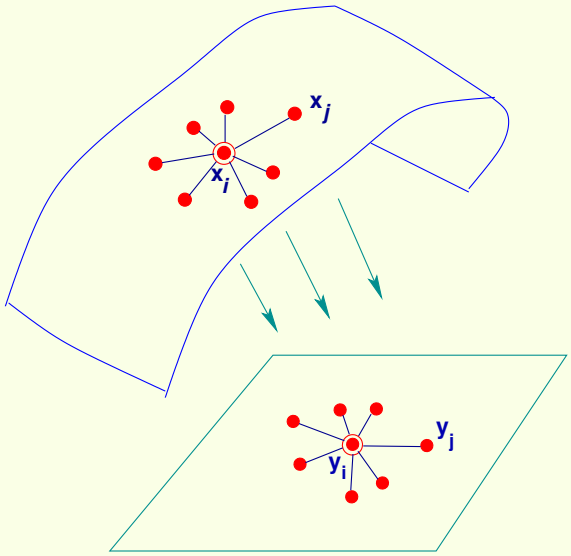
[3] S. Yan and X. O. Tang, "Trace ratio revisited" Proceedings of the European Conference on Computer Vision, 2006.

...

Graph-based methods

- Start with a graph of data. e.g.: graph of k nearest neighbors (k-NN graph)

Want: Perform a projection which preserves the graph in some sense



- Define a *graph Laplacean:*

$$L = D - W$$

e.g.,: $w_{ij} = \begin{cases} 1 & \text{if } j \in N_i \\ 0 & \text{else} \end{cases}$ $\quad D = \text{diag}\left[d_{ii} = \sum_{j \neq i} w_{ij} \right]$

with N_i = neighborhood of i (excluding i)

Example: The Laplacean eigenmaps approach

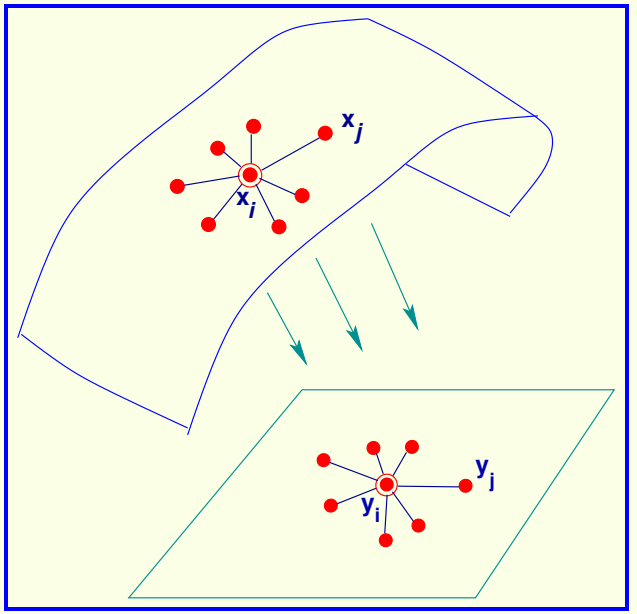
Laplacean Eigenmaps [Belkin-Niyogi '01] *minimizes*

$$\mathcal{F}(Y) = \sum_{i,j=1}^{n} w_{ij} \|y_i - y_j\|^2 \quad \text{subject to} \quad YDY^\top = I$$

Motivation: if $\|x_i - x_j\|$ is small (orig. data), we want $\|y_i - y_j\|$ to be also small (low-Dim. data)

➤ Original data used indirectly through its graph

➤ Leads to $n \times n$ sparse eigenvalue problem [In 'sample' space]



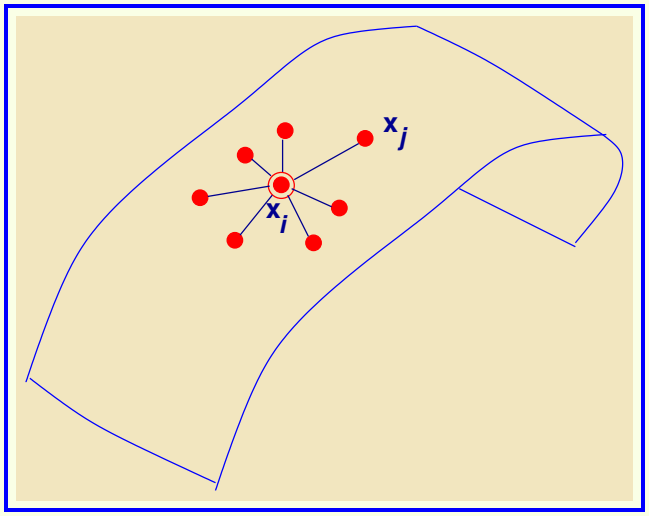
Locally Linear Embedding (Roweis-Saul-00)

- Very similar to Eigenmaps.
- Graph Laplacean matrix is replaced by an 'affinity' graph

Graph: Each x_i written as a convex combination of its k nearest neighbors:

$x_i \approx \Sigma w_{ij} x_j, \quad \sum_{j \in N_i} w_{ij} = 1$

- Optimal weights computed ('local calculation') by minimizing $\|x_i - \Sigma w_{ij} x_j\|$ for $i = 1, \cdots, n$



- Mapped data (Y) computed by minimizing

$$\sum \|y_i - \Sigma w_{ij} y_j\|^2$$

ONPP (Kokiopoulou and YS '05)

- Orthogonal Neighborhood Preserving Projections
- A linear (orthogonoal) version of LLE obtained by writing $\boldsymbol{Y}$ in the form $\boldsymbol{Y} = \boldsymbol{V}^\top \boldsymbol{X}$
- Same graph as LLE. Objective: preserve the affinity graph (as in LEE) *but* with the constraint $\boldsymbol{Y} = \boldsymbol{V}^\top \boldsymbol{X}$
- Problem solved to obtain mapping:

$$\min_{\boldsymbol{V}} \text{Tr}\left[\boldsymbol{V}^\top \boldsymbol{X}(\boldsymbol{I} - \boldsymbol{W}^\top)(\boldsymbol{I} - \boldsymbol{W})\boldsymbol{X}^\top \boldsymbol{V}\right]$$

$$\text{s.t. } \boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I}$$

- In LLE replace $\boldsymbol{V}^\top \boldsymbol{X}$ by $\boldsymbol{Y}$

Face Recognition – background

Problem: We are given a database of images: [arrays of pixel values]. And a test (new) image.

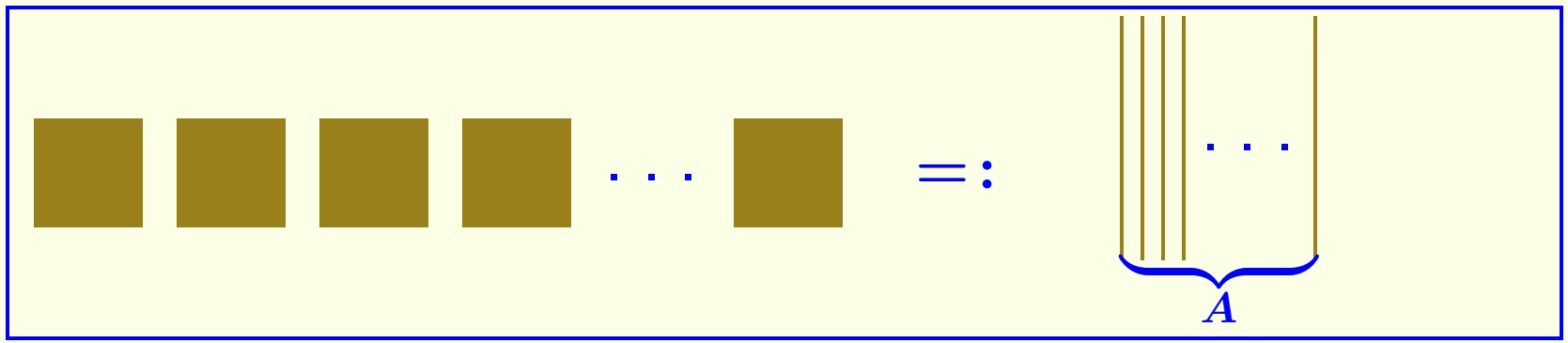
Question: Does this new image correspond to one of those in the database?

Example: Eigenfaces [Turk-Pentland, '91]

➤ Idea identical with the one we saw for digits:

– Consider each picture as a (1-D) column of all pixels

– Put together into an array A of size $\#\_pixels \times \#\_images$.



– Do an SVD of A and perform comparison with any test image in low-dim. space

– Similar to LSI in spirit – but data is not sparse.

Graph-based methods in a supervised setting

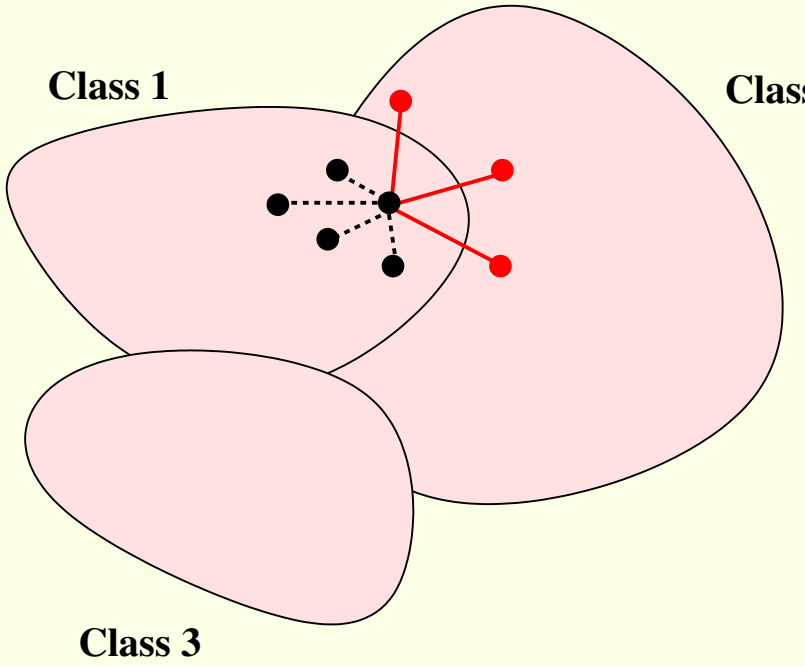
Graph-based methods can be adapted to supervised mode. Idea: Build G so that nodes in the same class are neighbors. If c = # classes, G consists of c cliques.

- Weight matrix W =block-diagonal
- Note: $\mathbf{rank}(W) = n - c$.
- As before, graph Laplacean:

$$L_c = D - W$$

$$W = \begin{pmatrix} W_1 & & & \\ & W_2 & & \\ & & \ddots & \\ & & & W_c \end{pmatrix}$$

- Can be used for ONPP and other graph based methods
- Improvement: add repulsion Laplacean [Kokiopoulou, YS 09]



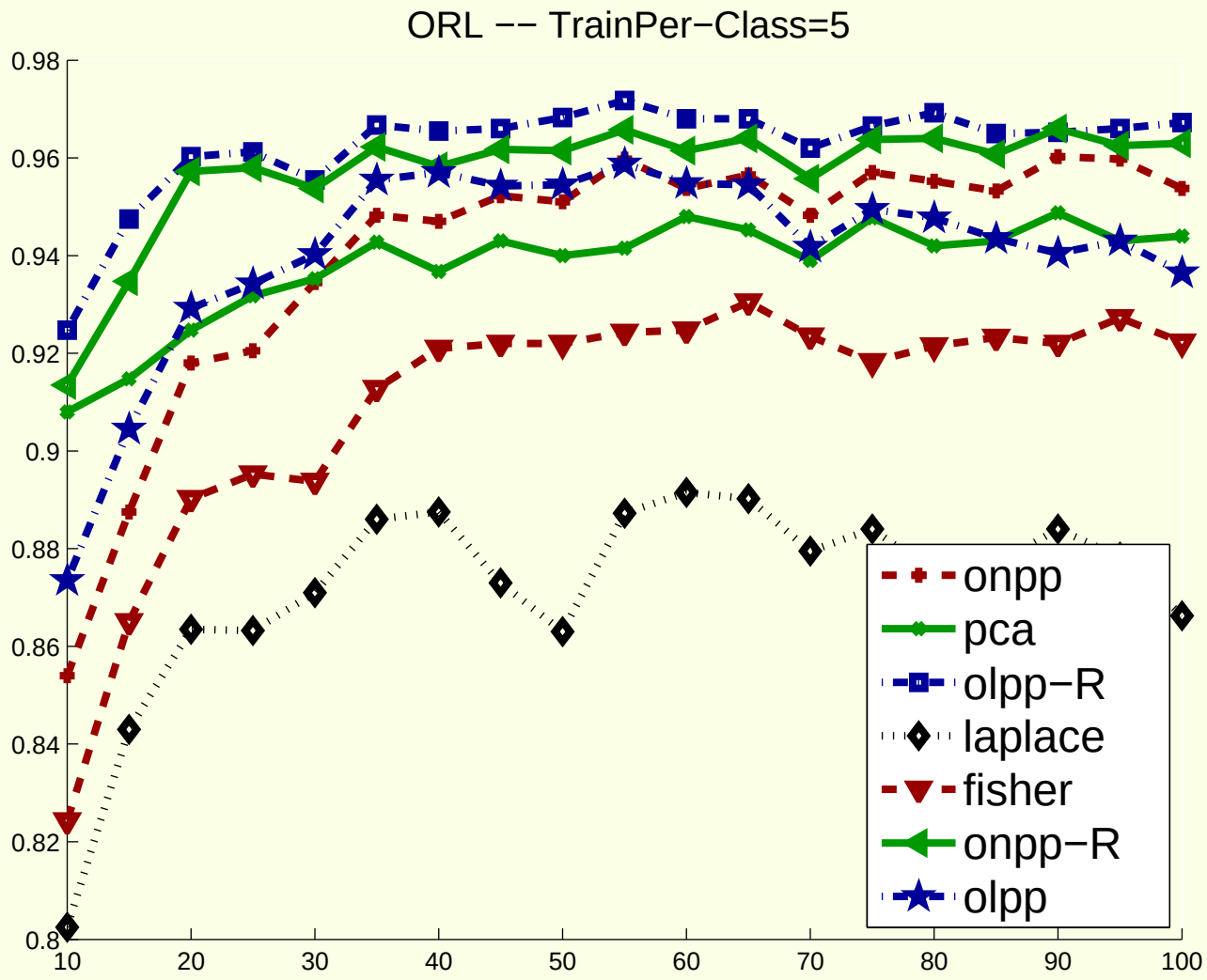
Leads to eigenvalue problem with matrix:

$$L_c - \rho L_R$$

- L_c = class-Laplacean,
- L_R = repulsion Laplacean,
- $\rho =$ parameter

Test: ORL 40 subjects, 10 sample images each – example:

# of pixels : 112×92; TOT. # images : 400



➤ Remarkable: there are values of ρ which give better results than using the optimum ρ obtained from maximizing trace ratio

Data mining for materials: Materials Informatics

*Collabor.: J. Chelikowsky (UT Austin), & Da Gao (U. of M)

➤ Huge potential in exploiting two trends:

1 Improvements in efficiency and capabilities in computational methods for materials

2 Recent progress in data mining techniques

➤ Current practice: "One student, one alloy, one PhD" [see special MRS issue on materials informatics] $\longrightarrow$ Slow ..

➤ Data Mining: can help speed-up process, e.g., by exploring in smarter ways

Issue 1: Who will do the work? Few researchers are familiar with both worlds

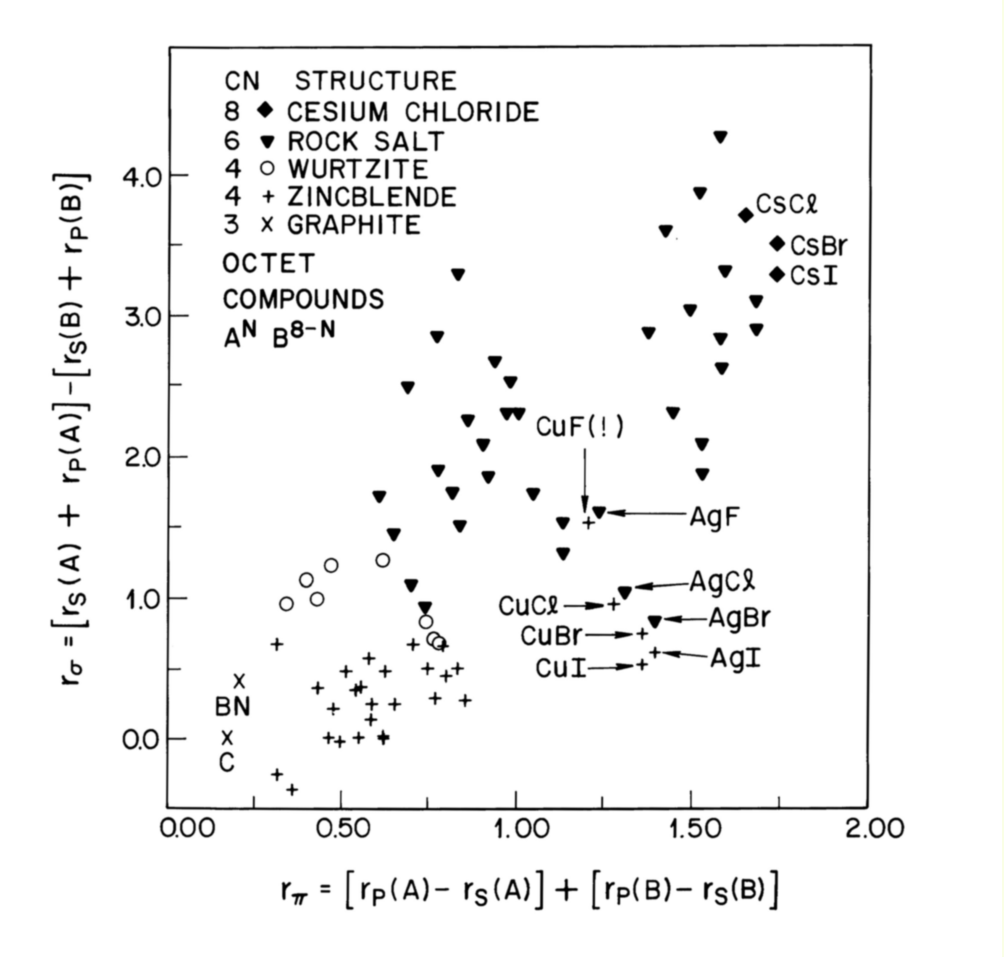
Issue 2: databases, and more generally sharing, not too common in materials

> *The inherently fragmented and multidisciplinary nature of the materials community poses barriers to establishing the required networks for sharing results and information. One of the largest challenges will be encouraging scientists to think of themselves not as individual researchers but as part of a powerful network collectively analyzing and using data generated by the larger community. These barriers must be overcome.*

NSTC report to the White House, June 2011.

- Materials genome initiative [NSF]

Unsupervised learning



- 1970s: Data Mining "by hand": Find coordinates that will cluster materials according to structure
- 2-D projection from physical knowledge
- 'Anomaly Detection': helped find that compound Cu F does not exist

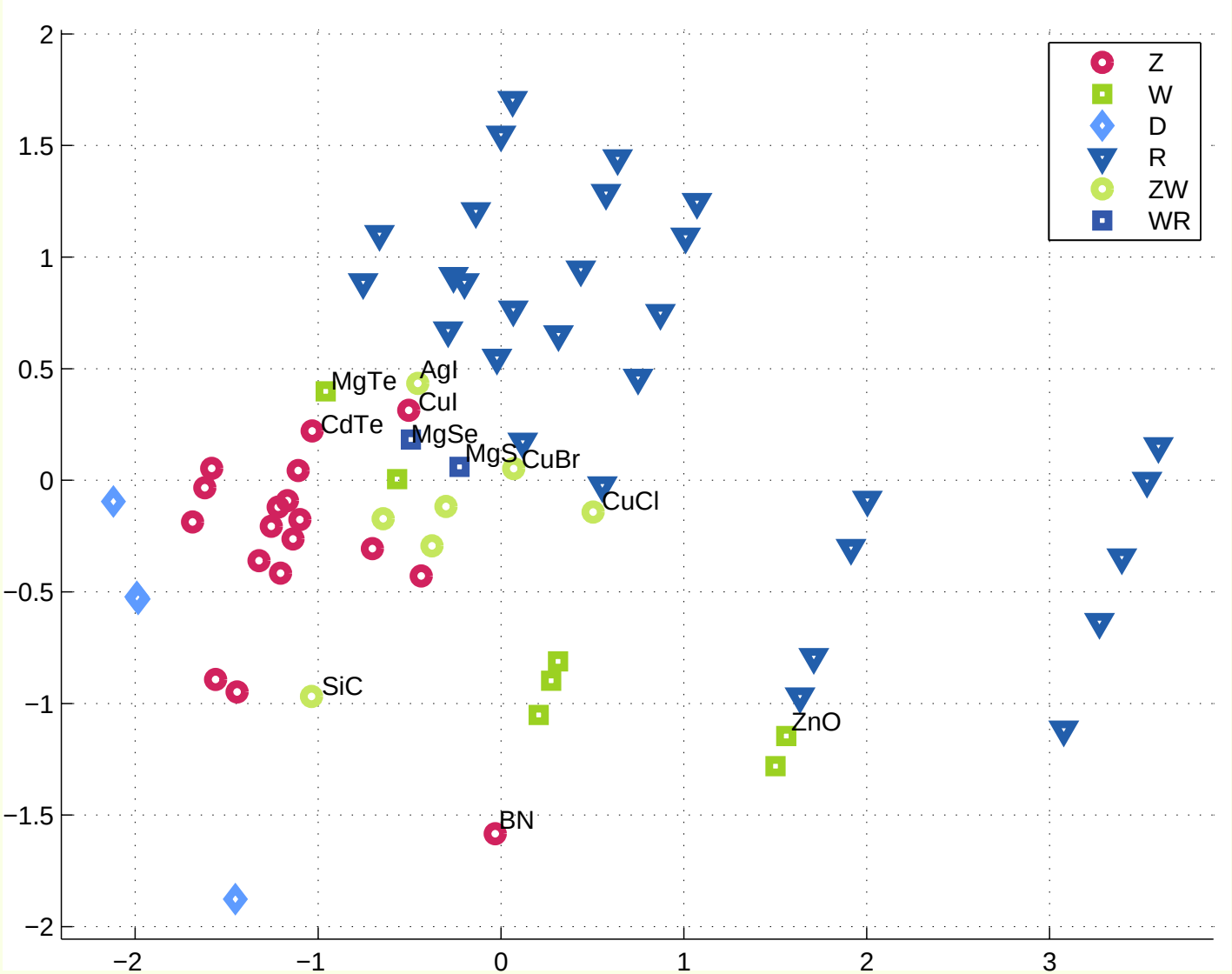
see: J. R. Chelikowsky, J. C. Phillips, Phys Rev. B 19 (1978).

Question: Can modern data mining achieve a similar diagrammatic separation of structures?

- Should use only information from the two constituent atoms
- Experiment: 67 binary 'octets'.
- Use PCA – exploit only data from 2 constituent atoms:

1. Number of valence electrons;
2. Ionization energies of the s-states of the ion core;
3. Ionization energies of the p-states of the ion core;
4. Radii for the s-states as determined from model potentials;
5. Radii for the p-states as determined from model potentials.

Result:


Z
W
D
R
ZW
WR
MgTe
AgI
CuI
CdTe
MgSe
MgS
CuBr
CuCl
SiC
ZnO
BN
2
1.5
1
0.5
0
−0.5
−1
−1.5
−2
−2
−1
0
1
2
3

Supervised learning: classification

- Problem: classify an unknown binary compound into its crystal structure class
- 55 compounds, 6 crystal structure classes
- "leave-one-out" experiment

Case 1: Use features 1:5 for atom A and 2:5 for atom B. No scaling is applied.

Case 2: Features 2:5 from each atom + scale features 2 to 4 by square root of # valence electrons (feature 1)

Case 3: Features 1:5 for atom A and 2:5 for atom B. Scale features 2 and 3 by square root of # valence electrons.

Three methods tested

1. PCA classification. Project and do identification in space of reduced dimension (Euclidean distance in low-dim space).

2. KNN K-nearest neighbor classification –

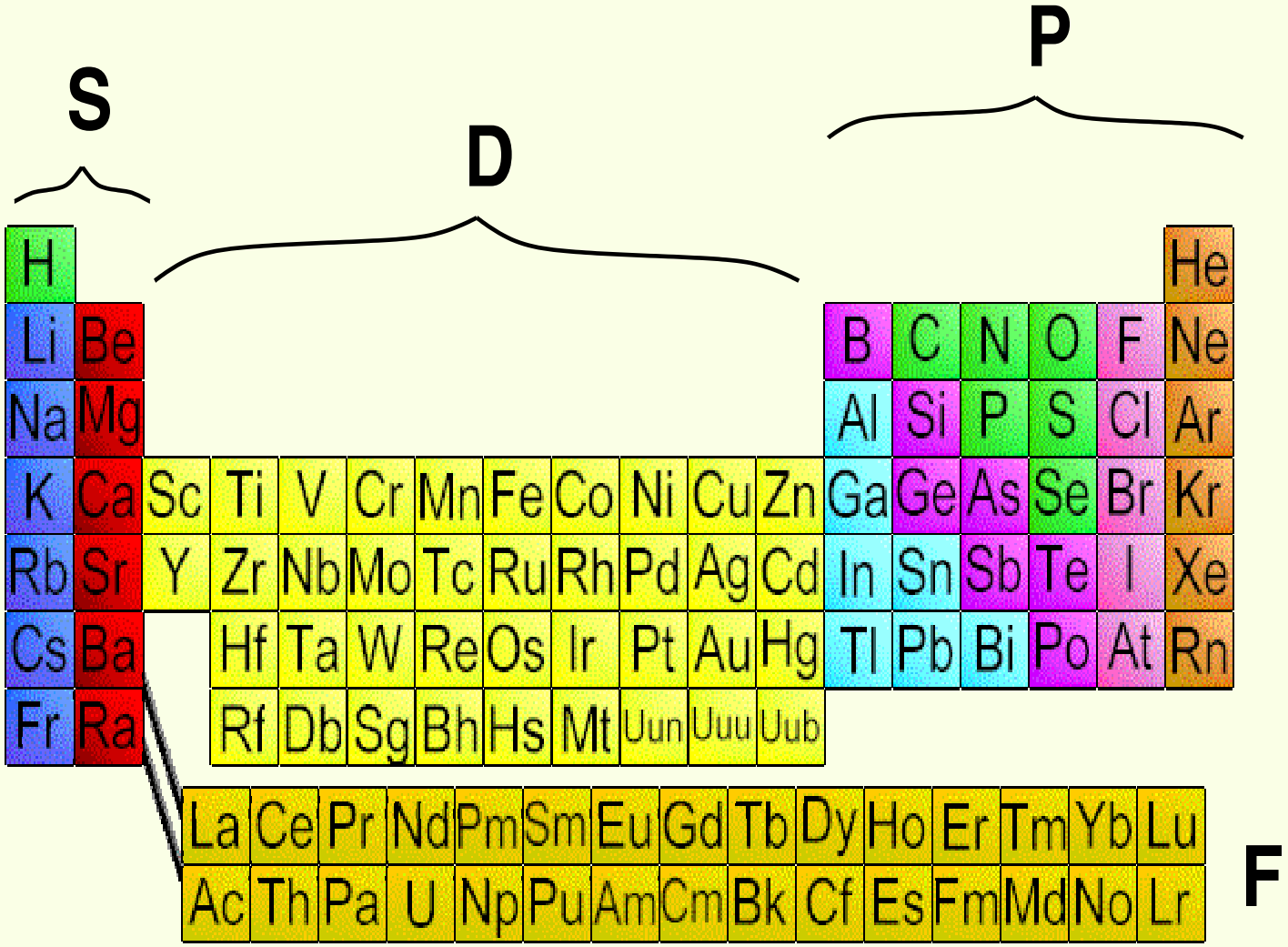
3. Orthogonal Neighborhood Preserving Projection (ONPP) - a graph based method - [see Kokiopoulou, YS, 2005]

Recognition rates for 3 different methods using different features

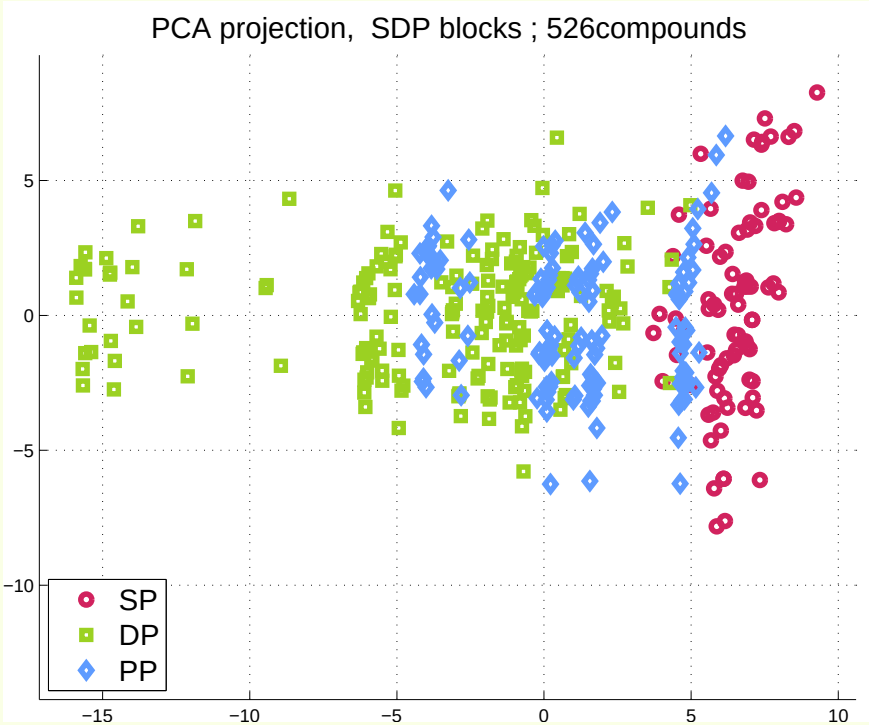
Case	KNN	ONPP	PCA
Case 1	0.909	0.945	0.945
Case 2	0.945	0.945	1.000
Case 3	0.945	0.945	0.982

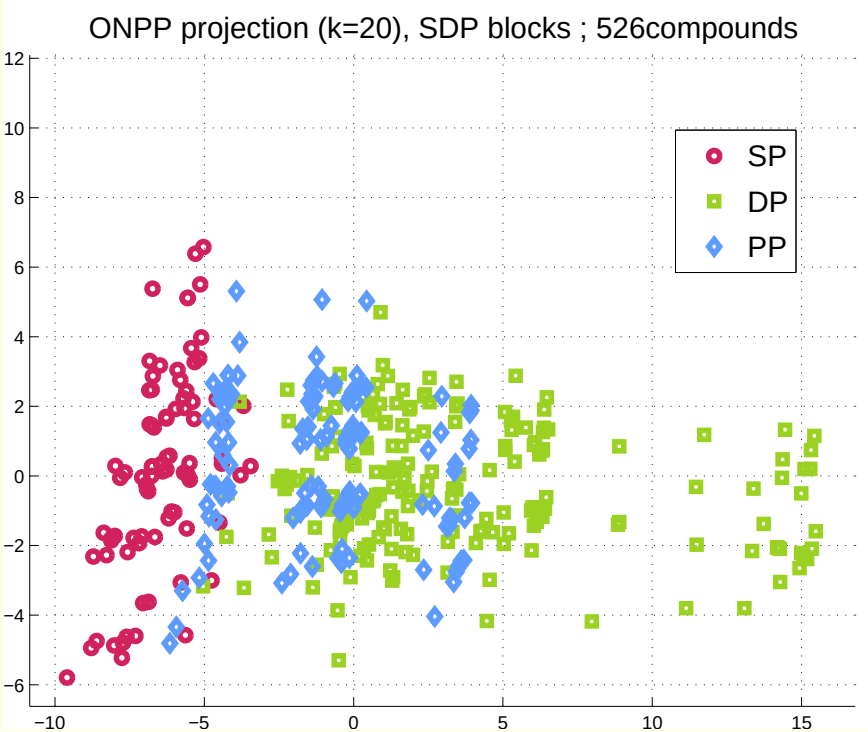
A few experiments with a larger database

- 529 A_xB_y compounds - Extract 'valid' entries
- Source: Semiconductors, Data Handbook Otfried Madelung, Springer; 3rd edition (January 22, 2004)
- Band-gap information available -
- Experiments : [unsupervised learning]
  - 2-D projection using SPDF blocks
  - 2-D projection showing Insulator - Semi-conductor property


S
D
P
H
He
Li Be
B C N O F Ne
Na Mg
Al Si P S Cl Ar
K Ca Sc Ti V Cr Mn Fe Co Ni Cu Zn Ga Ge As Se Br Kr
Rb Sr Y Zr Nb Mo Tc Ru Rh Pd Ag Cd In Sn Sb Te I Xe
Cs Ba Hf Ta W Re Os Ir Pt Au Hg Tl Pb Bi Po At Rn
Fr Ra Rf Db Sg Bh Hs Mt Uun Uuu Uub
La Ce Pr Nd Pm Sm Eu Gd Tb Dy Ho Er Tm Yb Lu
Ac Th Pa U Np Pu Am Cm Bk Cf Es Fm Md No Lr
F

Results for SDP 2-D visualization

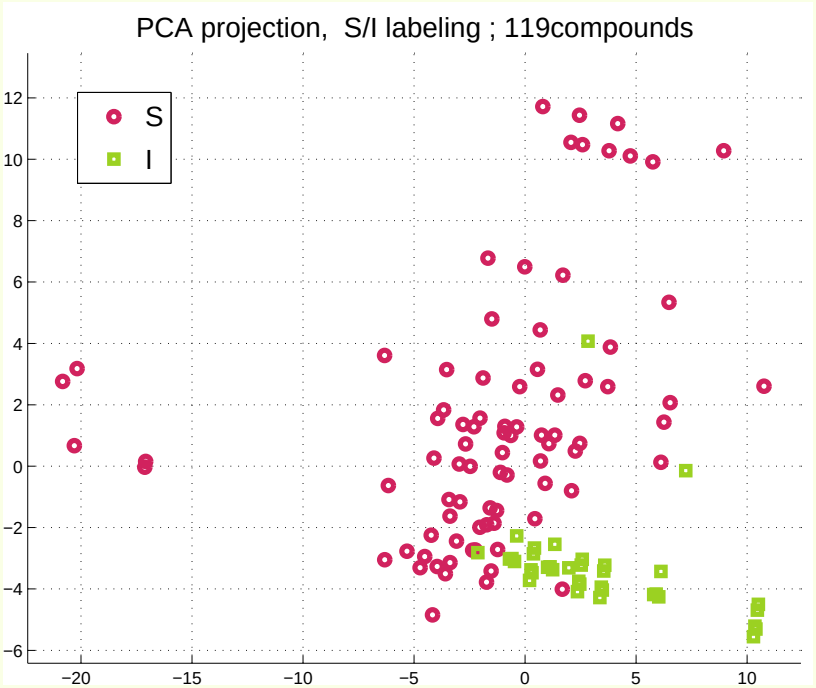


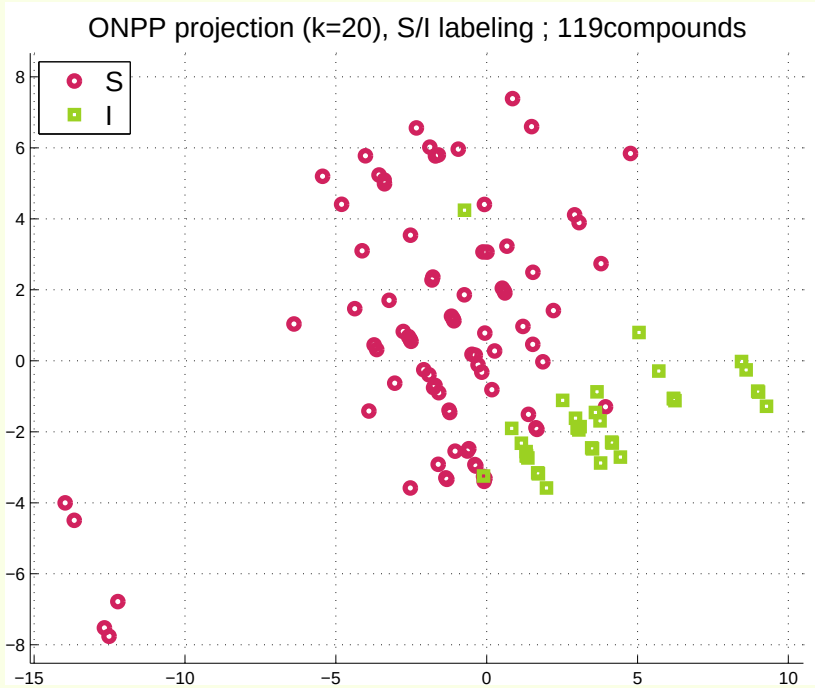


Features used: val. elec.; electronegativ.; covalent rad.; density; ionic rad.; chemical scale; + band-gap.

Results for 2-D Semiconductor/Insulator visualization

- Had to consider AB only compounds [119 of them]





features used: val. electrons; electro-negativity; cov. rad.; density; ionic rad.;

Conclusion

- Many, interesting new matrix problems in areas that involve the effective mining of data
- Among the most pressing issues is that of reducing computational cost - [SVD, SDP, ... too costly]
- Many online resources available in computer science/ applied math...
- .. but there is an inertia to overcome. In particular ...
- ... data-sharing is not as widespread in materials science
- Plus side: Materials informatics will likely be energized by the materials genome project.

When one door closes, another opens; but we often look so long and so regretfully upon the closed door that we do not see the one which has opened for us.

Alexander Graham Bell (1847-1922)

Thank you !

- Visit my web-site at `www.cs.umn.edu/~saad`