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

ICSC 2012, Hong Kong, Jan 4-7, 2012

HAPPY BIRTHDAY TONY!

Introduction: What is data mining?

➤ Common goal of data mining methods: to extract meaningful information or patterns from data. Very broad area – includes: data analysis, machine learning, pattern recognition, information retrieval, ...

➤ Main tools used: linear algebra; graph theory; approximation theory; statistics; optimization; ...

➤ In this talk: brief introduction with emphasis on dimension reduction; Applications.

Major tool of Data Mining: Dimension reduction

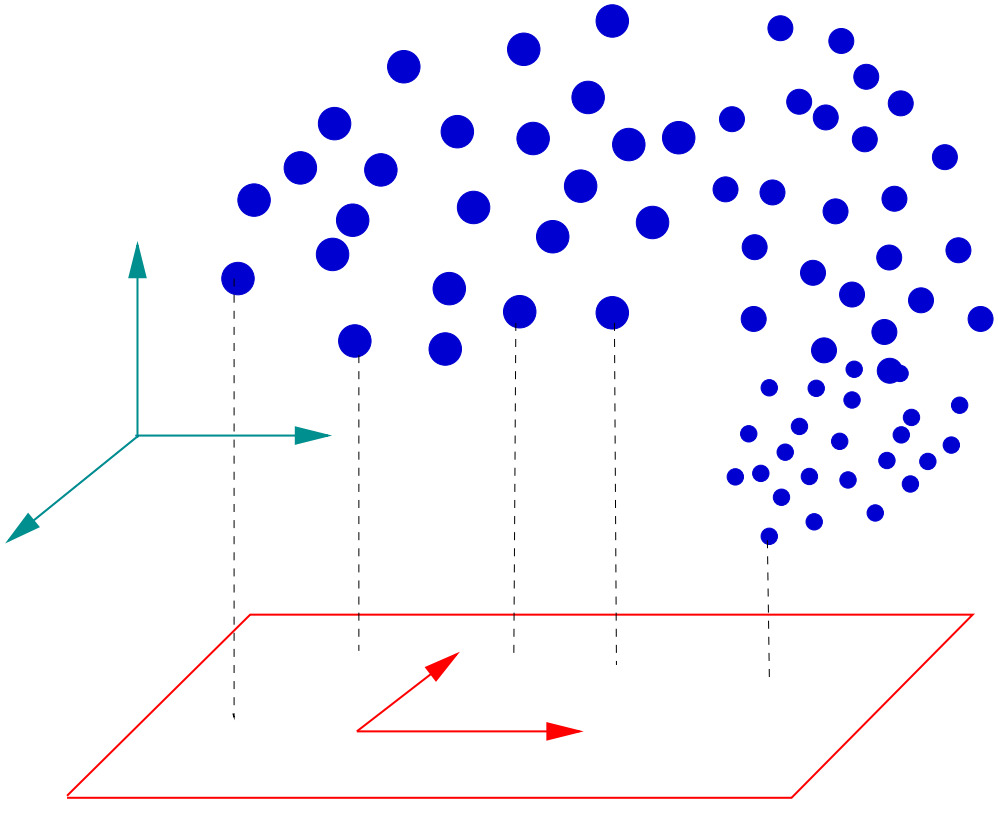
- Given $d \ll m$ find a mapping

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

- Mapping may be explicit [typically linear], e.g.: $y = V^T x$
- Or implicit (nonlinear)
- Techniques depend on application: Preserve angles? Preserve distances? Maximize variance? ..
- Primary goals of dimension reduction :
  - Reduce noise and redundancy in data
  - Discover 'features' or 'patterns'

Basic linear dimensionality reduction: PCA

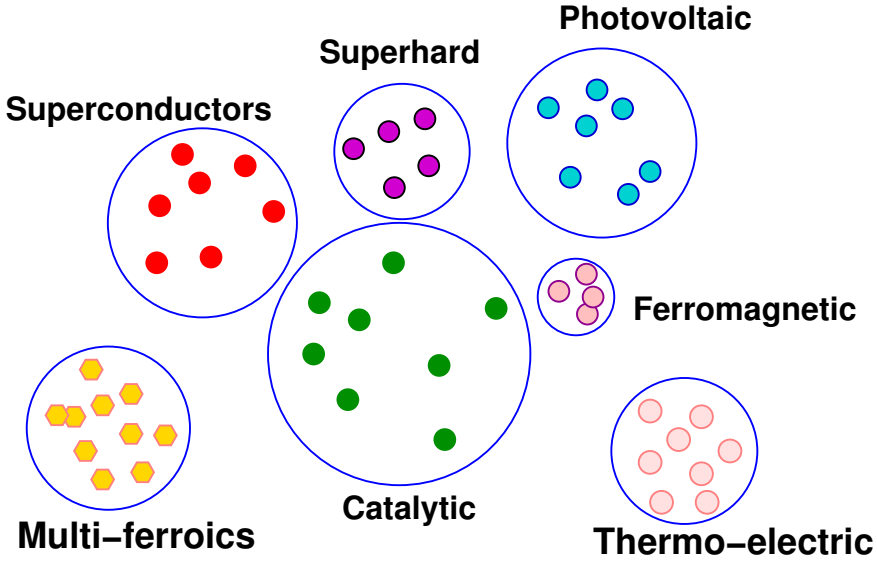
- We are given points in $\mathbb{R}^n$ and we want to project them in $\mathbb{R}^d$. Best way to do this?
- i.e.: find the best axes for projecting the data
- Q: "best in what sense"?
- A: maximize variance of new data



- Principal Component Analysis [PCA]

Unsupervised learning: Clustering

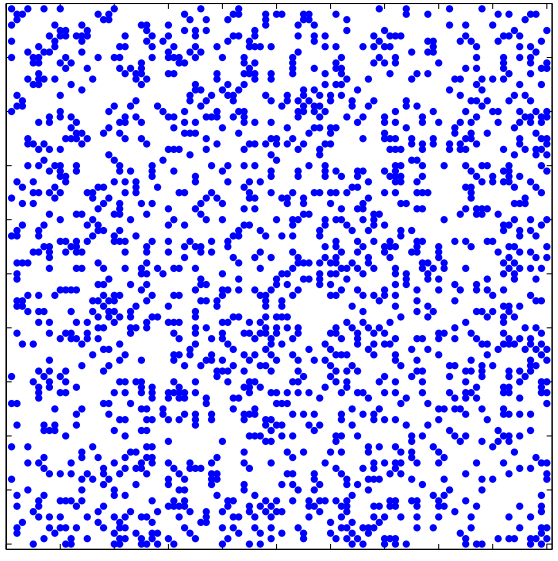
Problem: partition a given set into subsets such that items of the same subset are most similar and those of two different subsets most dissimilar.



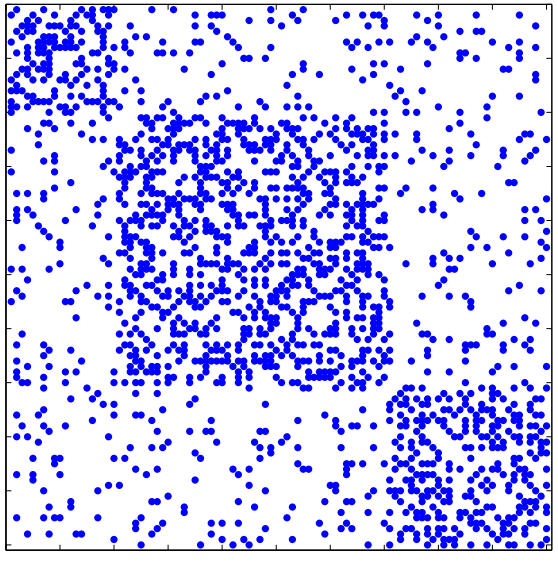
- Basic technique: K-means algorithm [slow but effective.]
- Example of application : cluster bloggers by 'social groups'

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



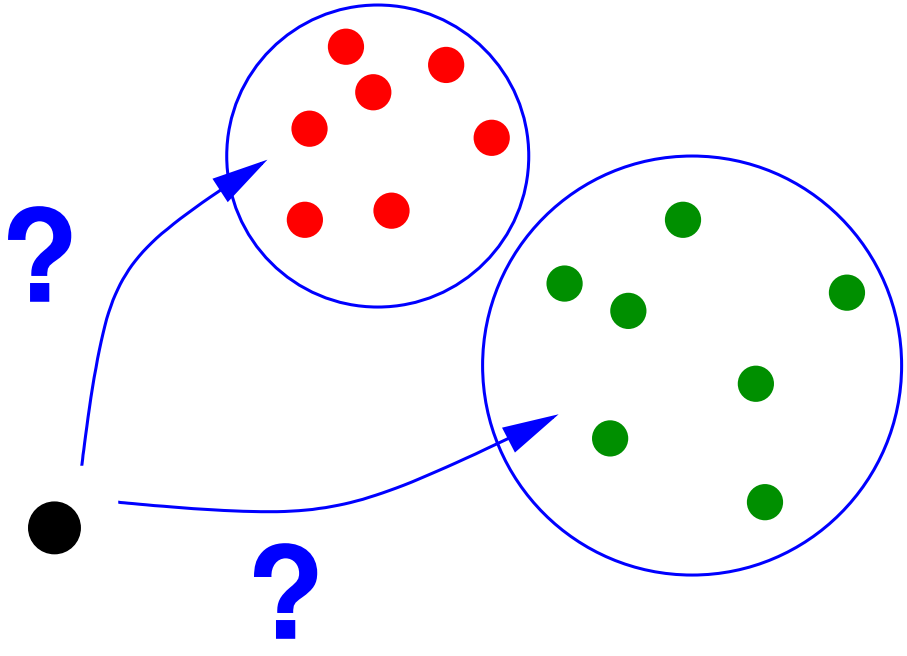
- Goal: find ordering so blocks are as dense as possible:



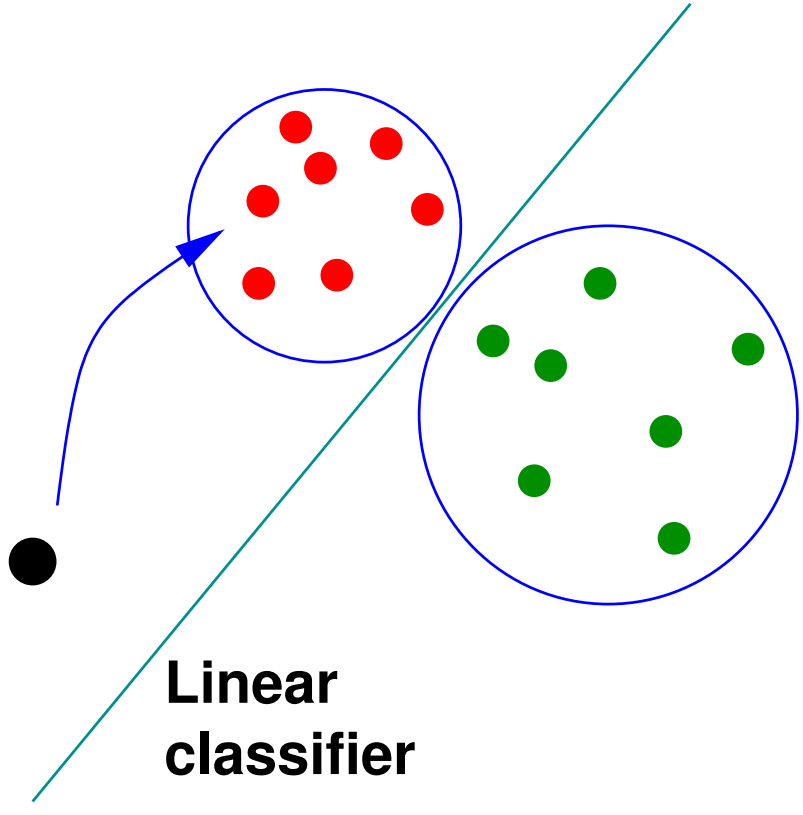
- Advantage of this viewpoint: need not know # of clusters.
- Use 'blocking' techniques for sparse matrices

Supervised learning: classification

Problem: Given labels (say "A" and "B") for each item of a given set, find a mechanism to classify an unlabelled item into either the "A" or the "B" class.



- Many applications.
- Example: distinguish SPAM and non-SPAM messages
- Can be extended to more than 2 classes.



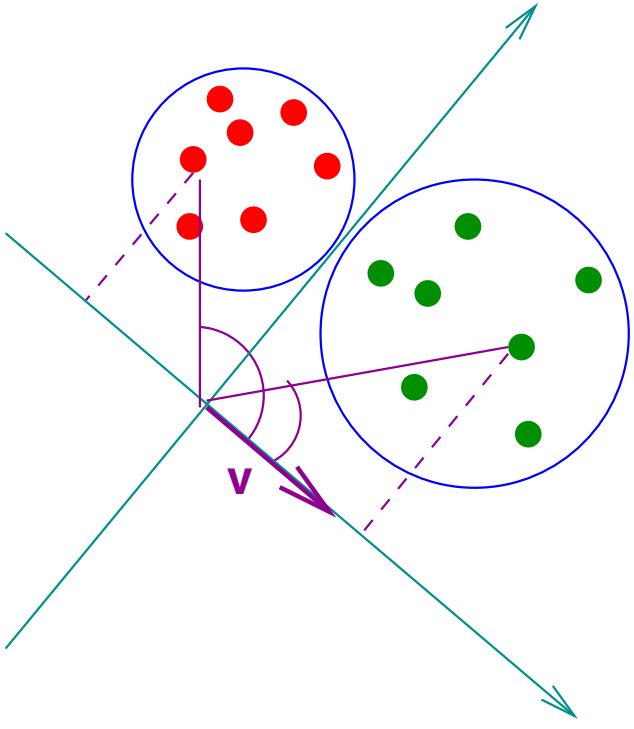
Linear classifiers: Find a hyperplane which best separates the data in classes A and B.

Linear classifiers

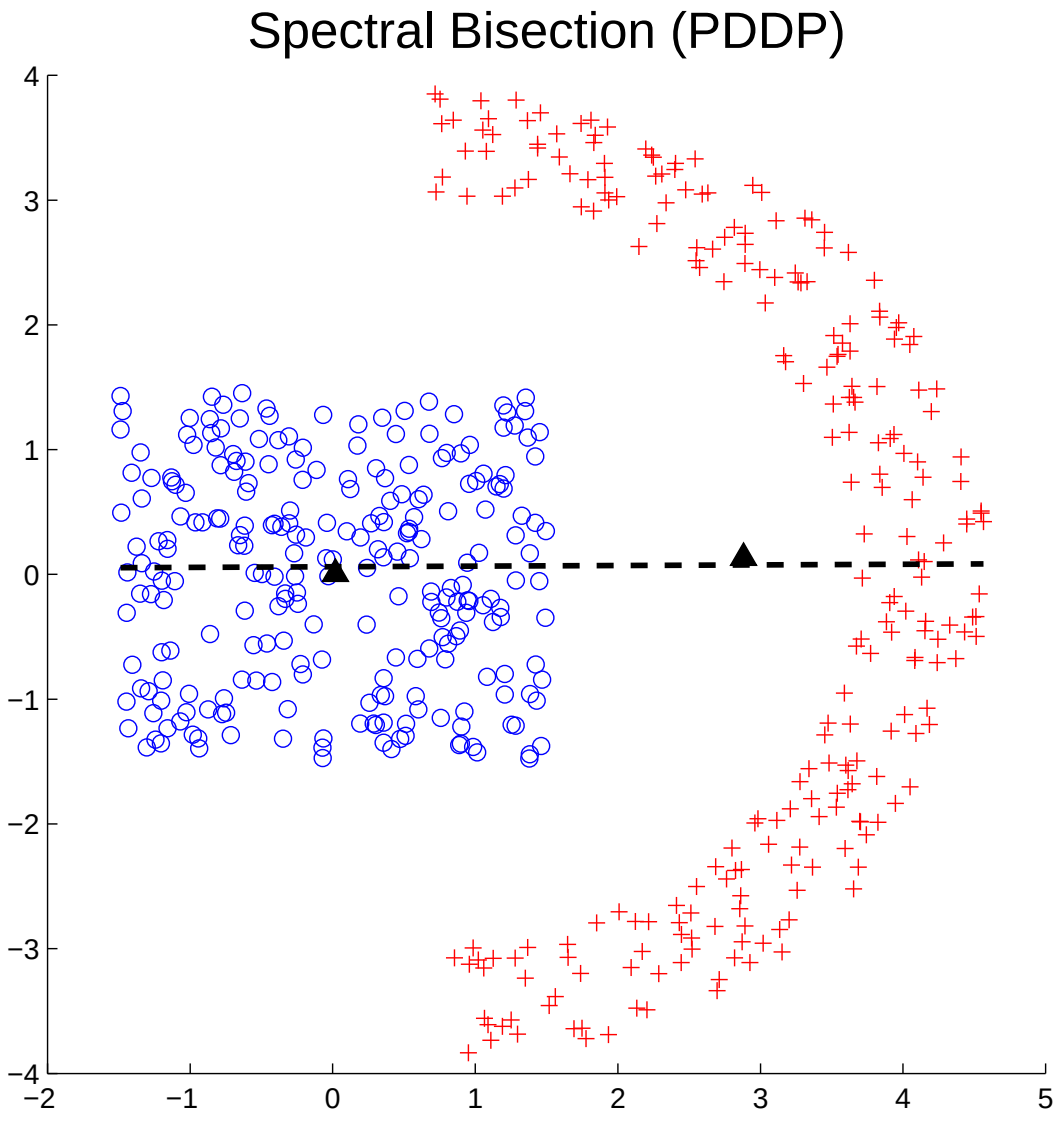
- Let $\boldsymbol{X} = [\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n]$ be the data matrix.
- and $\boldsymbol{L} = [\boldsymbol{l}_1, \cdots, \boldsymbol{l}_n]$ the labels either +1 or -1.

- 1st Solution: Find a vector $\boldsymbol{v}$ such that $\boldsymbol{v}^T \boldsymbol{x}_i$ close to $\boldsymbol{l}_i \; \forall \boldsymbol{i}$
- Common solution: SVD to reduce dimension of data [e.g. 2-D] then do comparison in this space. e.g.

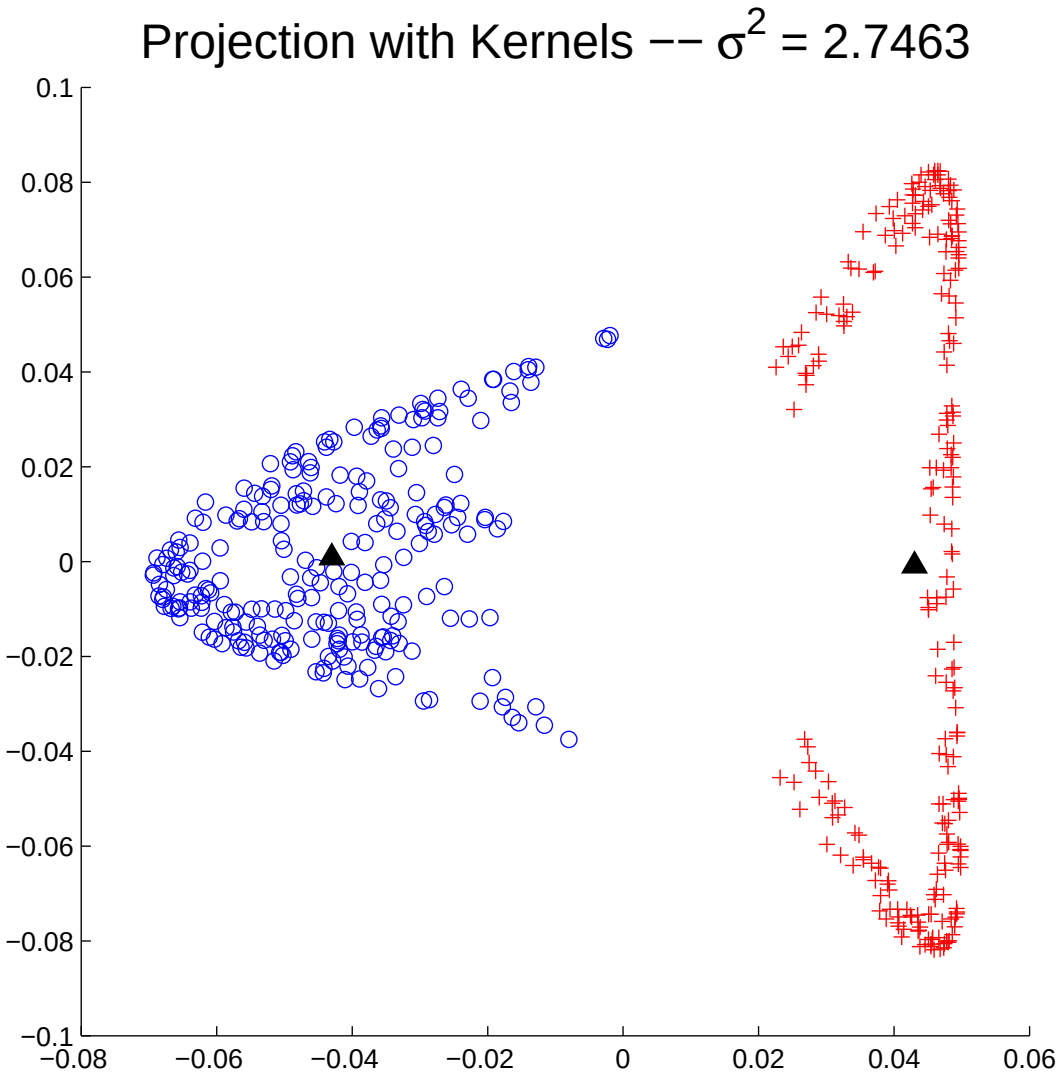
A: $\boldsymbol{v}^T \boldsymbol{x}_i \geq \mathbf{0}$, B: $\boldsymbol{v}^T \boldsymbol{x}_i < \mathbf{0}$



Need for nonlinear classifiers



- Result not too good. Use kernels to transform



Transformed data with a Gaussian Kernel

Linear Discriminant Analysis (LDA)

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.

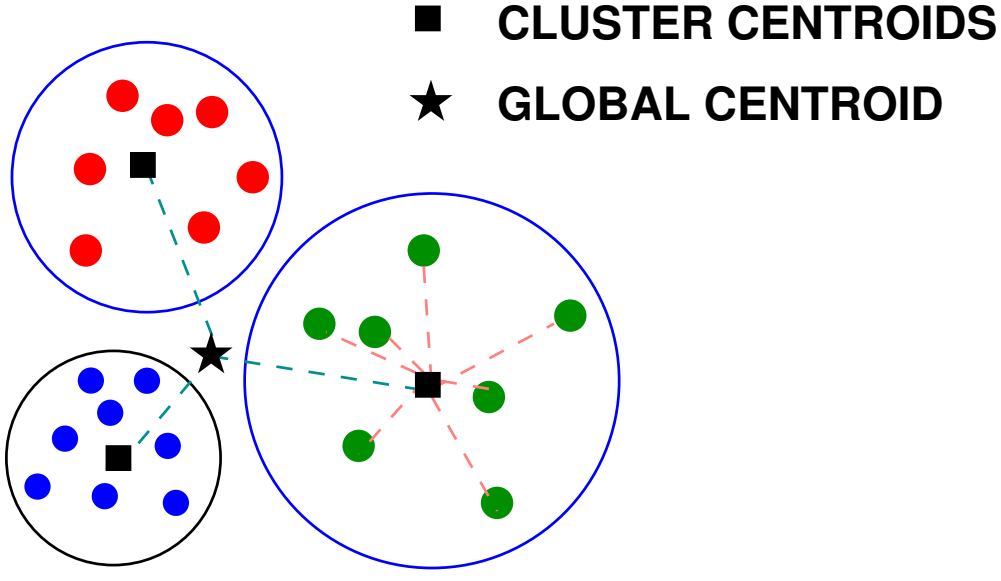
➤ Goal: to make "between scatter" measure large, while making "within scatter" small.

Idea: Project the data in low-dimensional space so as to maximize the ratio of the "between scatter" measure over "within scatter" measure of the classes.

Let μ = mean of X, and $\mu^{(k)}$ = mean of the k-th class (of size n_k). Define:

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

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



➤ Project set on a one-dimensional space spanned by a vector a.
Then:

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

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

➤ 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 dimension

- Would like to project in d dimensions –
- Wish to maximize the ratio of two traces s.t. $U^T U = I$

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

- Reduced dimension data: $Y = U^T X$.

Common belief: Hard to maximize. In fact not a big issue – See Ngo, Bellalij & YS

- Common alternative: Solve instead the (easier) problem:

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

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

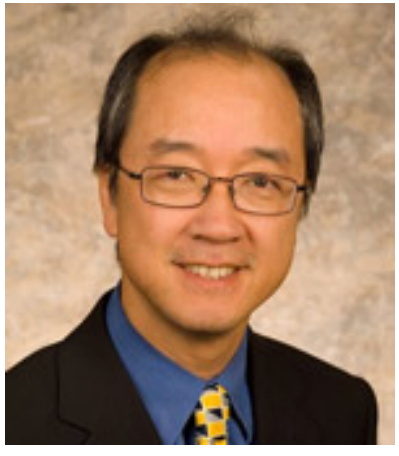
Face Recognition – background

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?

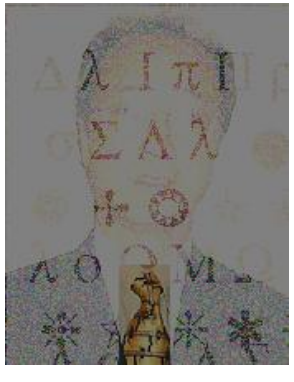
Difficulty Positions, Expressions, Background, Lighting, Neck-ties, ...,



Eigenfaces: Principal Component Analysis technique

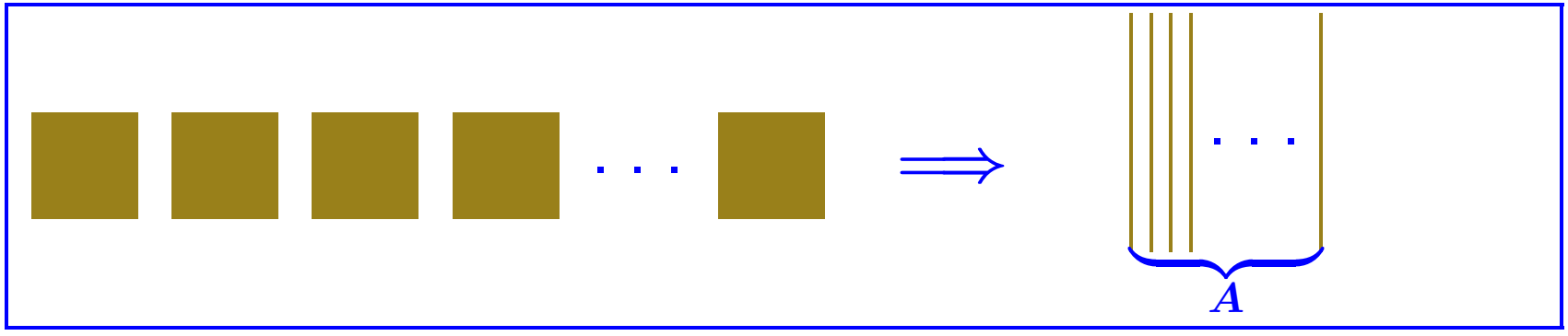
- Specific situation: Poor images or deliberately altered images ['occlusion']
- See real-life examples – [international man-hunt]



Eigenfaces

– 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.

Idea: replace SVD by Lanczos vectors (same as for IR)

Tests with two well-known data sets

ORL 40 subjects, 10 sample images each – example:

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

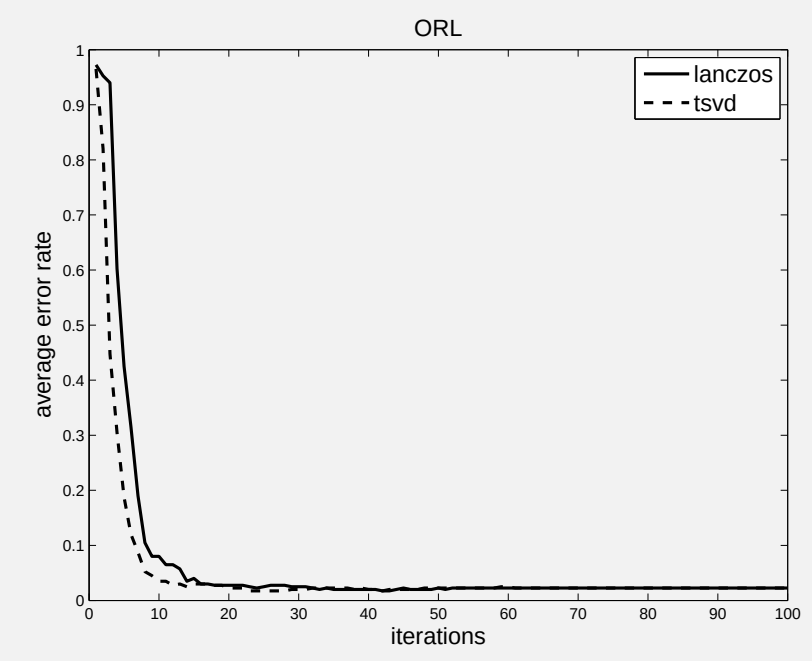
AR set 126 subjects – 4 facial expressions selected for each [natural, smiling, angry, screaming] – example:

# of pixels : 112×92; TOT. # images : 504

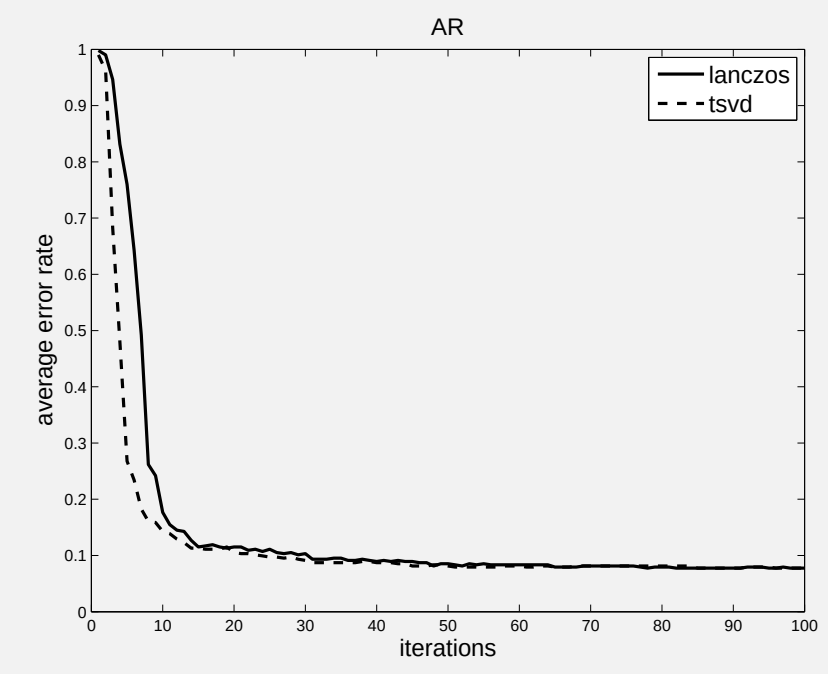
Tests: Face Recognition

Recognition accuracy of Lanczos approximation vs SVD

ORL dataset



AR dataset



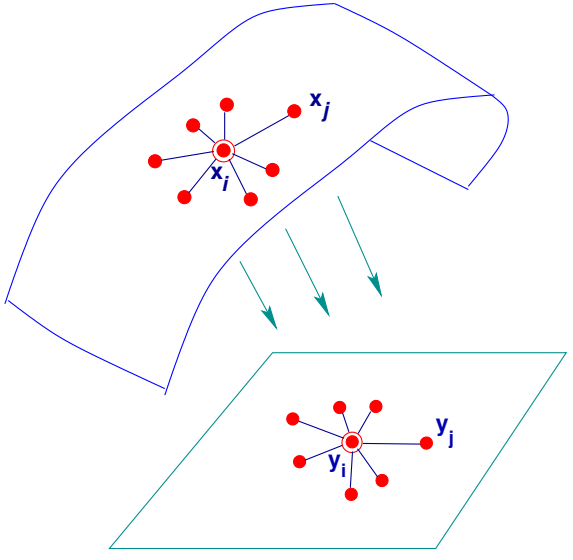
y-axis: average error rate. x-axis: Subspace dimension

GRAPH-BASED TECHNIQUES

Graph-based methods

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

Want: Do a projection so as to preserve the graph in some sense



➤ Define a *graph Laplacean:*

$$L = D - W$$

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

with N_i = neighborhood of i (excl. i)

Example: The Laplacean eigenmaps approach

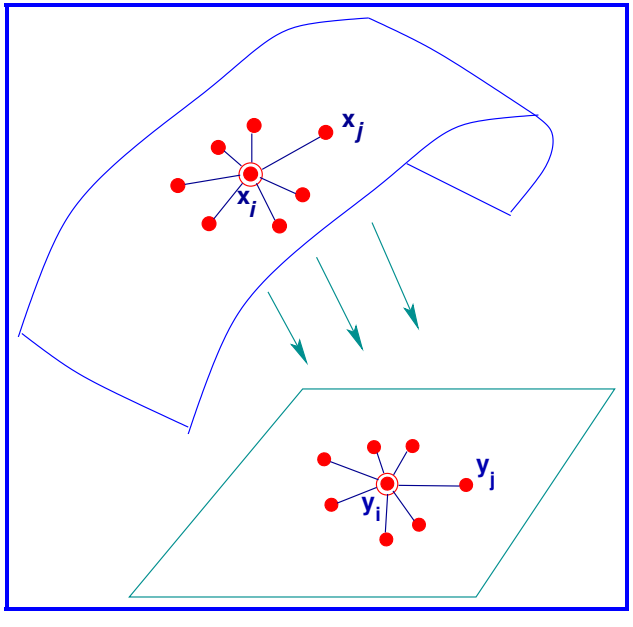
Laplacean Eigenmaps *minimizes*

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

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

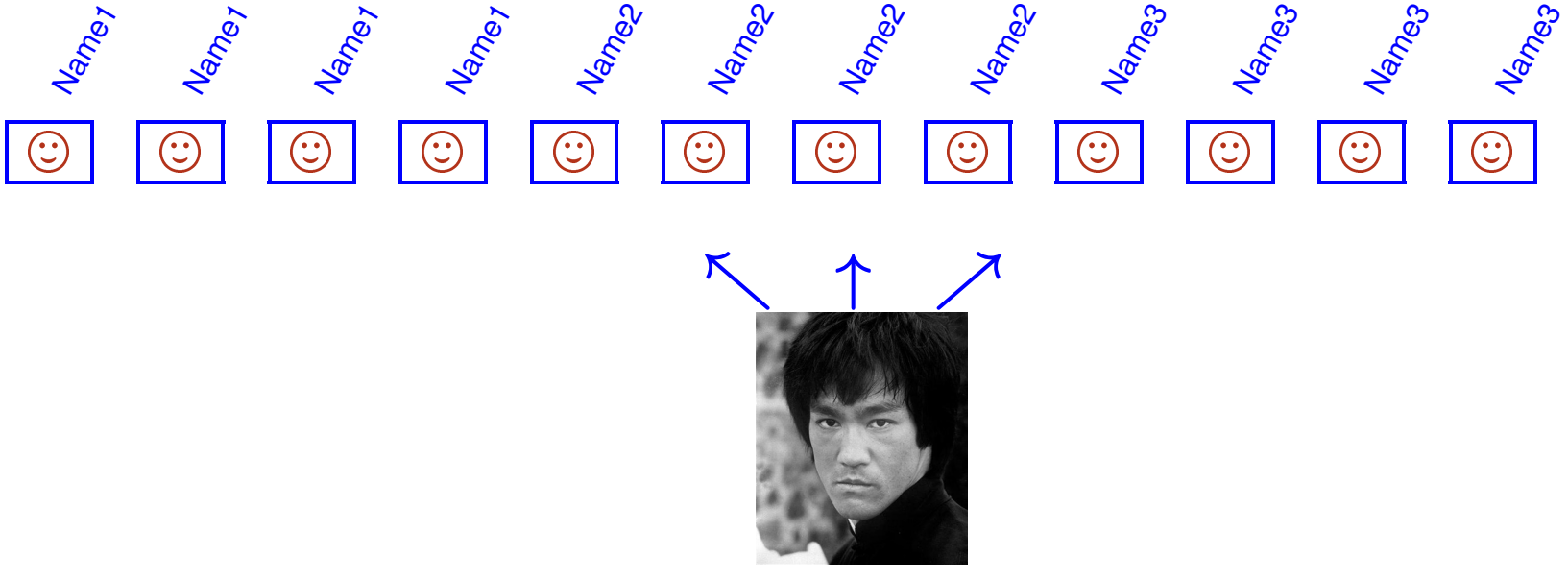
Note: Min instead of Max as in PCA

- Above problem uses original data indirectly through its graph
- Leads to $n \times n$ sparse eigenvalue problem [In 'sample' space]



Graph-based methods in a supervised setting

➤ Subjects of training set are known (labeled). Q: given a test image (say) find its label.



Question: Find label (best match) for test image.

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.

➤ Matrix W is block-diagonal

➤ Note: $\mathbf{rank}(W) = n - c$.

$$W = \begin{pmatrix} W_1 & & & & \\ & W_2 & & & \\ & & W_3 & & \\ & & & W_4 & \\ & & & & W_5 \end{pmatrix}$$

➤ Can be used for LPP, ONPP, etc..

➤ Recent improvement: add repulsion Laplacean [Kokiopoulou, YS 09]

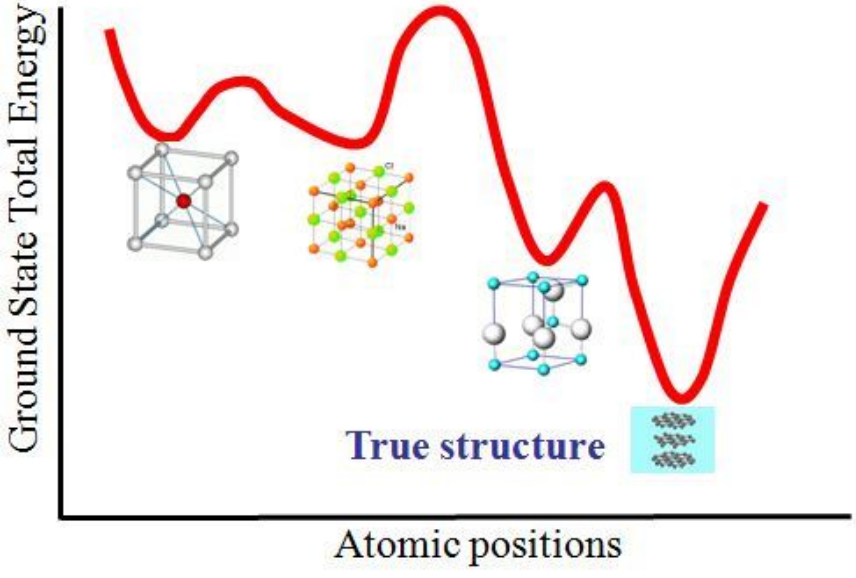
APPLICATION TO MATERIALS

Application to materials

➤ Studying materials properties using ab-initio methods can a major challenge

➤ Density Functional Theory & Kohn Sham equation used to determine electronic structure

➤ Often solved many many times for a single simulation



Kohn-Sham equations → nonlinear eigenvalue Pb

$$\left[-\frac{1}{2}\nabla^2 + (V_{ion} + V_H + V_{xc})\right]\Psi_i = E_i\Psi_i, i = 1, ..., n_o$$

$$\rho(r) = \sum_i^{n_o} |\Psi_i(r)|^2$$

$$\nabla^2 V_H = -4\pi\rho(r)$$

- Both V_{xc} and V_H, depend on ρ.
- Potentials & charge densities must be self-consistent.
- Most time-consuming part: diagonalization

Data mining for materials: Materials Informatics

➤ 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" → Slow pace of discovery

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

➤ However: databases, and generally sharing, in materials are not in great shape.

> *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.*

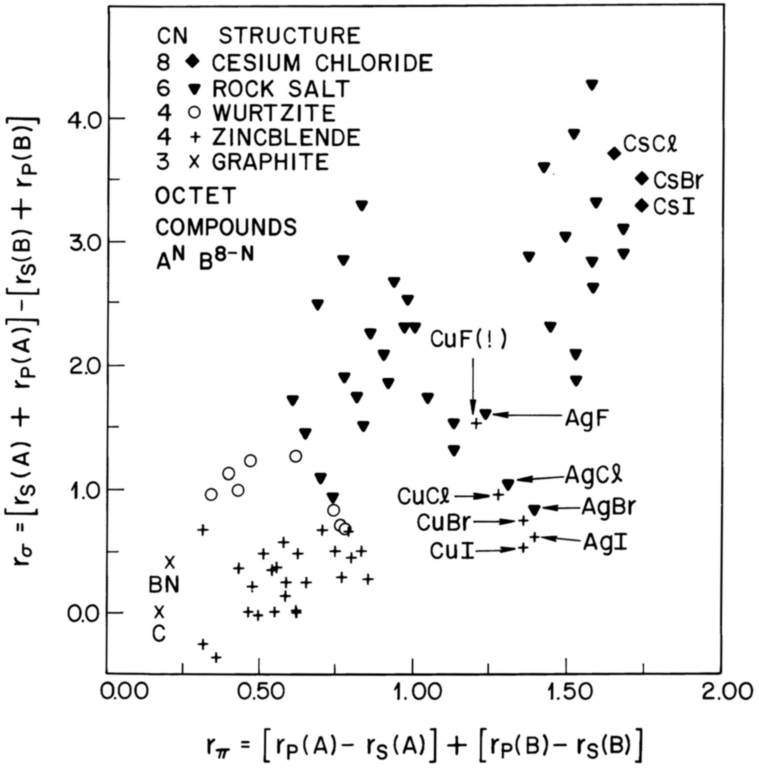
NSTC report to the white house, June 2011.

➤ Materials genome initiative [NSF]

Unsupervised clustering

➤ 1970s: Data Mining "by hand": Find coordinates to cluster materials according to structure

➤ 2-D projection from physical knowledge



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

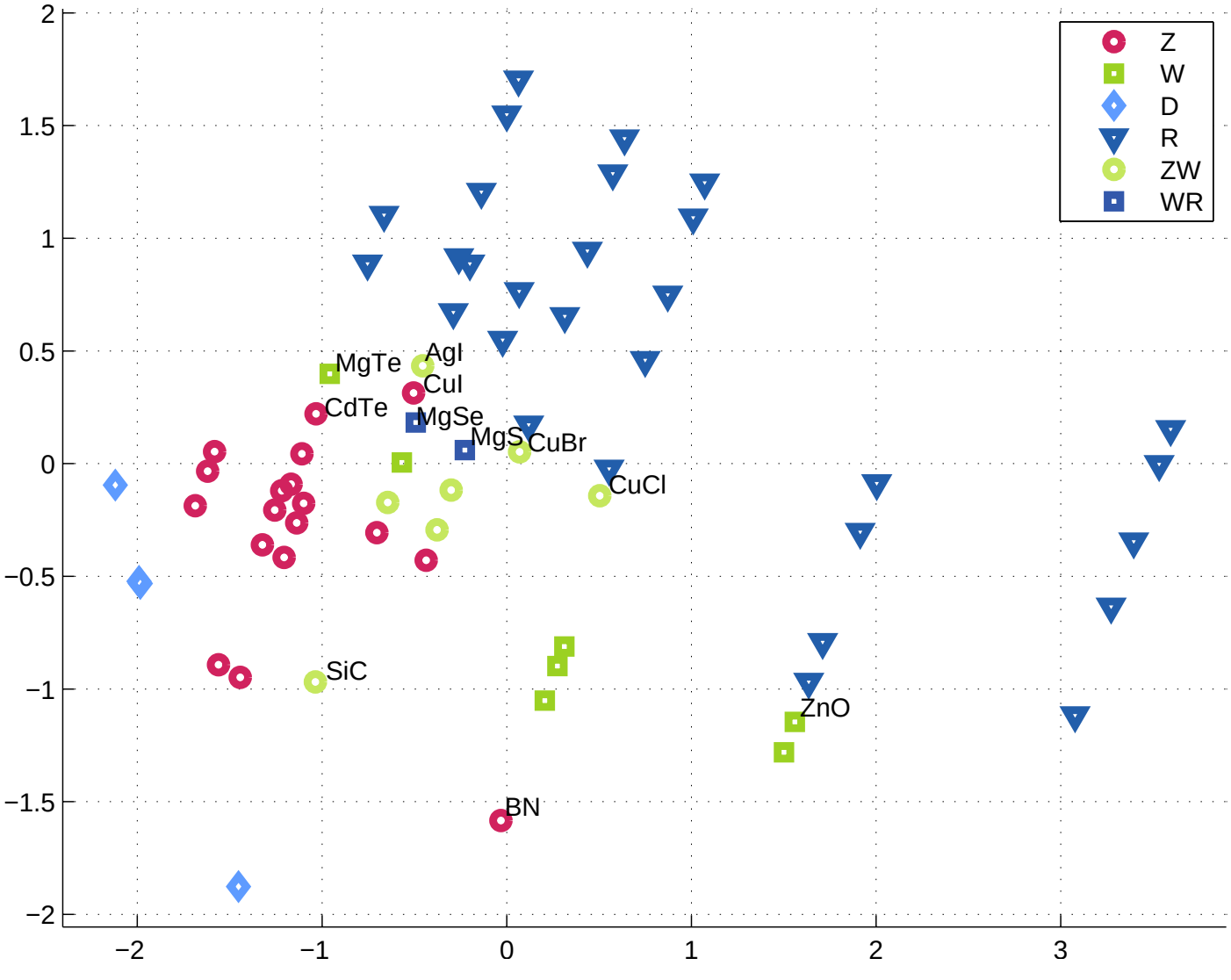
➤ 'Anomaly Detection': helped find that compound Cu F does not exist

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 –

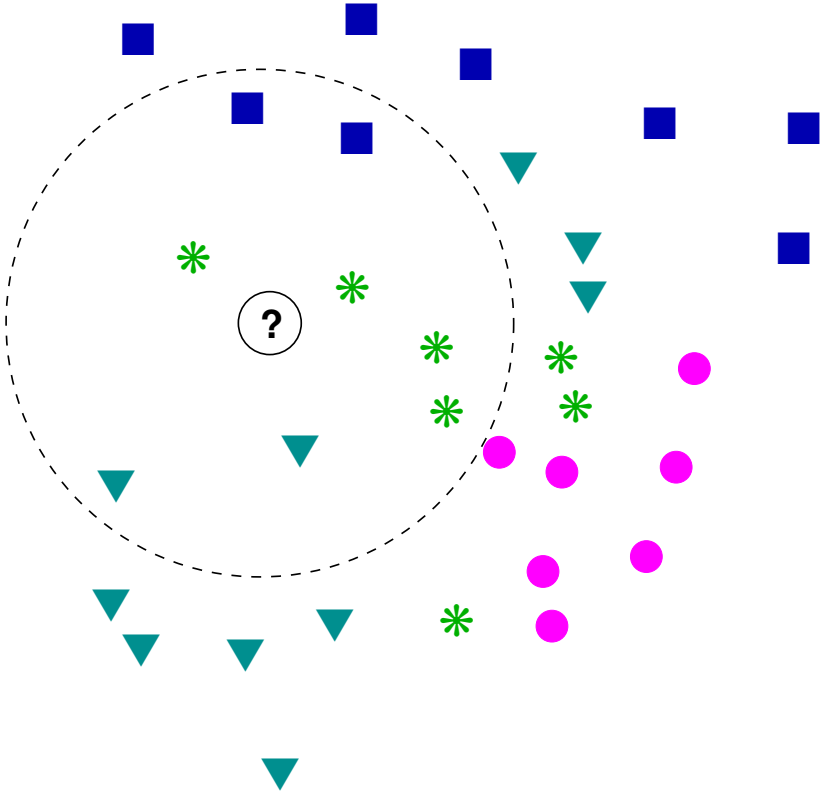
3. Orthogonal Neighborhood Preserving Projection (ONPP) - a graph based method - [see Kokiopoulou, YS, 2005]

K-nearest neighbor (KNN) classification

- Arguably the simplest of all methods
- Idea of a voting system:

get distances between test sample and all other compounds

- Classes of the k nearest neighbors are considered (k=8)
- The predominant class among these k items is assigned to the test sample ("asterisk" here)



Results

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

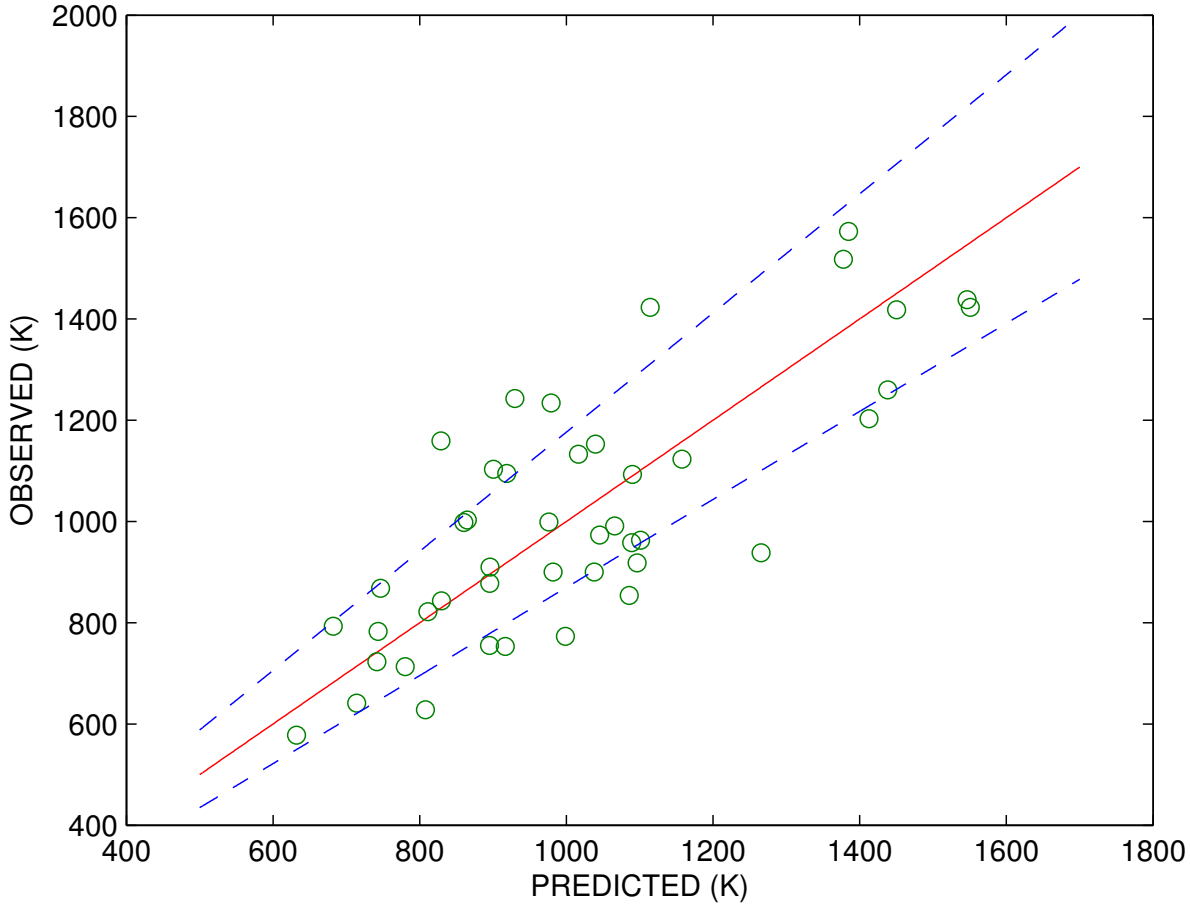
Recognition rate for 3 different methods using the data in different ways

Melting point prediction

- It can be very difficult to predict materials properties by correlations alone.
- Test: 44 AB suboctet compounds
- Experimental melting points used
- “Leave-one-out” experiment
- Use simple linear regression with Tikhonov regulation

Atomic features used:

(1) Number of valence electrons;
(2) Radius for s states as determined from model potentials;
(3) Radius for p states as determined from model potentials;
(4) Electron negativity;
(5) Boiling point;
(6) 1st ionization potential;
(7) Heat of vaporization;
(8) Atomic number.



Experimental and predicted melting points in degrees K. Blue dashed lines == boundaries for 15% relative error.

Conclusion

➤ Many, interesting new matrix problems related to data mining as well as emerging scientific fields:

1. Information technologies [learning, data-mining, ...]
2. Computational Chemistry / materials science
3. Bio-informatics: computational biology, genomics, ..

➤ *Important:* Lots of resources available online: repositories, tutorials,.. Easy to get started.

➤ Materials informatics: likely be energized by the materials genome project.