Supervised Distance Metric Learning

A Retrospective

Nan Xiao
Stat. Dept., Central South Univ.
Q4 2013

Outline

- Theory
- Algorithms
- Applications

What is a metric?

A metric is a function.

$$X_1 \times X_2 \to \mathbb{R}^1$$

Function D satisfies:

- $D(\mathbf{a}, \mathbf{b}) \geq 0$
- $D(\mathbf{a}, \mathbf{b}) = 0$, iff. $\mathbf{a} = \mathbf{b}$
- $D(\mathbf{a}, \mathbf{b}) = D(\mathbf{b}, \mathbf{a})$
- $D(\mathbf{a}, \mathbf{b}) + D(\mathbf{b}, \mathbf{c}) \geq D(\mathbf{a}, \mathbf{c})$

A metric is a similarity measure.

$$d_{\mathbf{M}}(x_i, x_j) = (x_i - x_j)^T \mathbf{M} (x_i - x_j)$$

Euclidean Distance

Minkowski Distance

Mahalanobis Distance

Levenshtein (Edit) Distance

...

Kernel Function

Distance Metrics as an Essential Building Block in PR & ML

- Unsupervised Learning (K-means)
- Supervised Learning (k-NN)

Distance Metrics as an Essential Building Block in PR & ML

- Computer Vision
- Natural Language Processing
- Information Retrieval (Text, Picture, Music, Video)
- Bioinformatics (DNA / Protein Sequence Alignment)

Metric Selection

- Priori
  - Euclidean Distance / Cosine Distance
  - Gaussian Kernel / Linear Kernel
- Experiment
  - Cross Validation

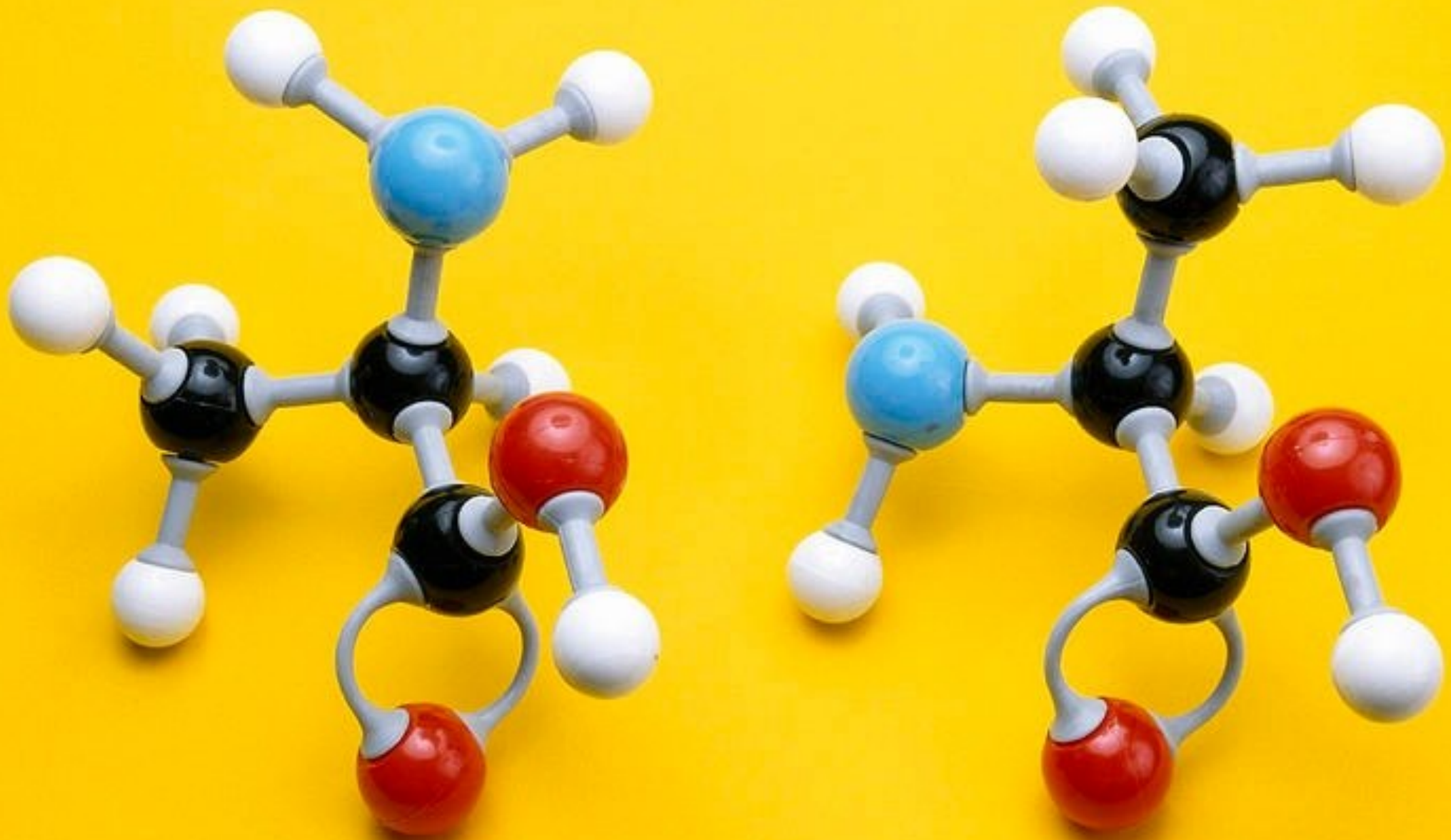
Why we need customized distance metrics?






Cougar Face Category from Caltech101 Dataset

Frome, A et al., (2007).Image Retrieval and Classification Using Local Distance Functions.



α-Alanine / β-Alanine

Similarity Baseline Effect in Large Scale Chemical DB

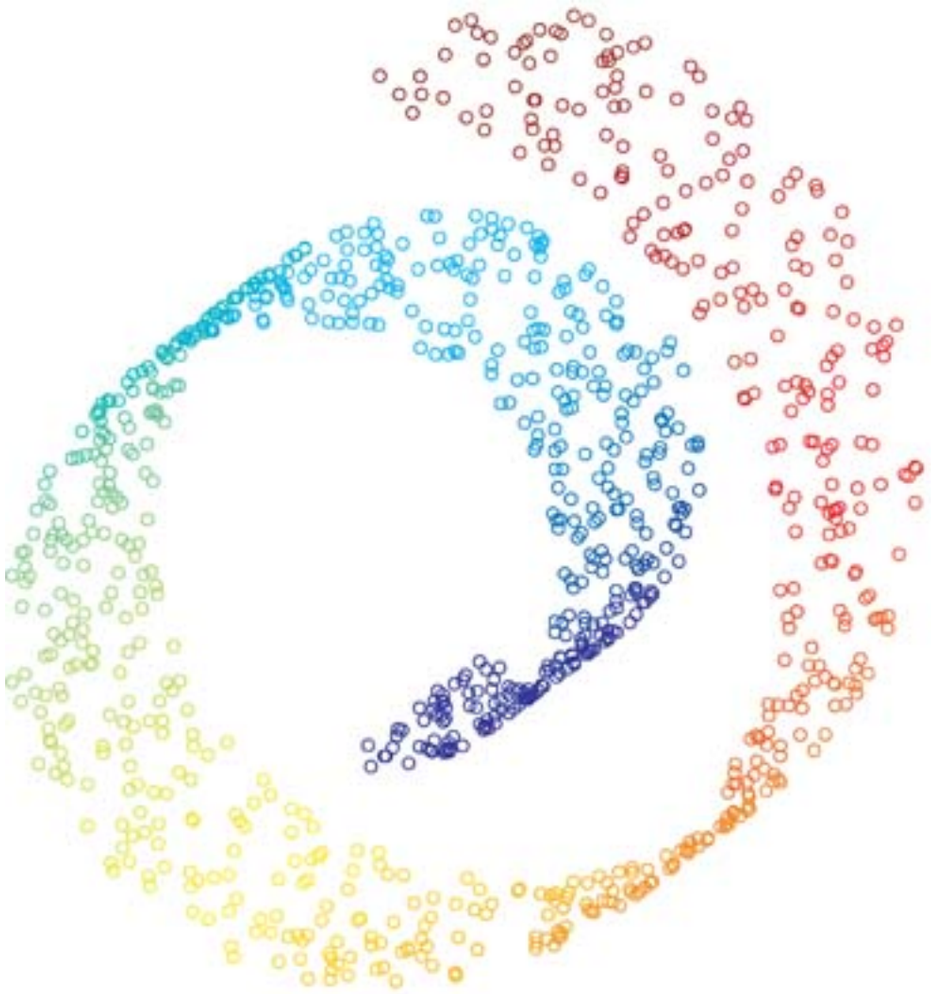
Rulers vs. Metrics

The Unsupervised Approach

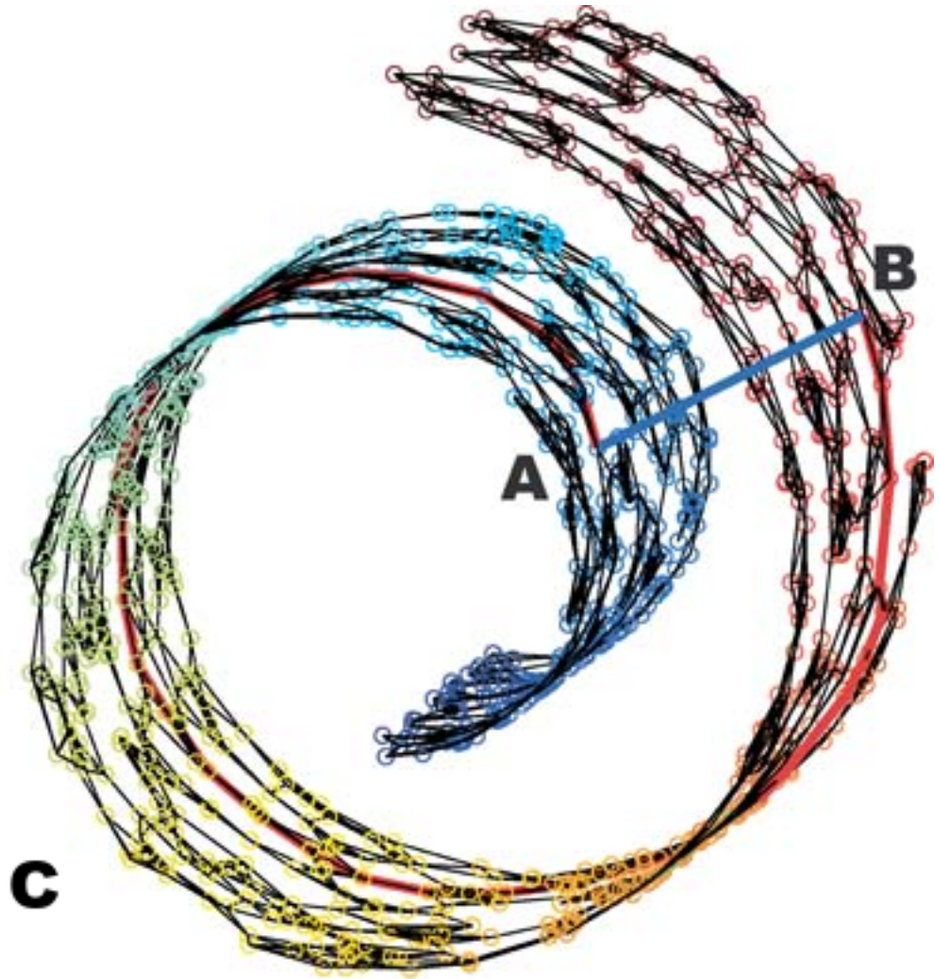
Manifold Learning

- PCA / MDS (Not Essentially)
- Isomap (Science, 2000)
- LLE / Hessian LLE
- LTSA
- Spectral Embedding
- and many more variants.

Real Swiss Roll …



(a)



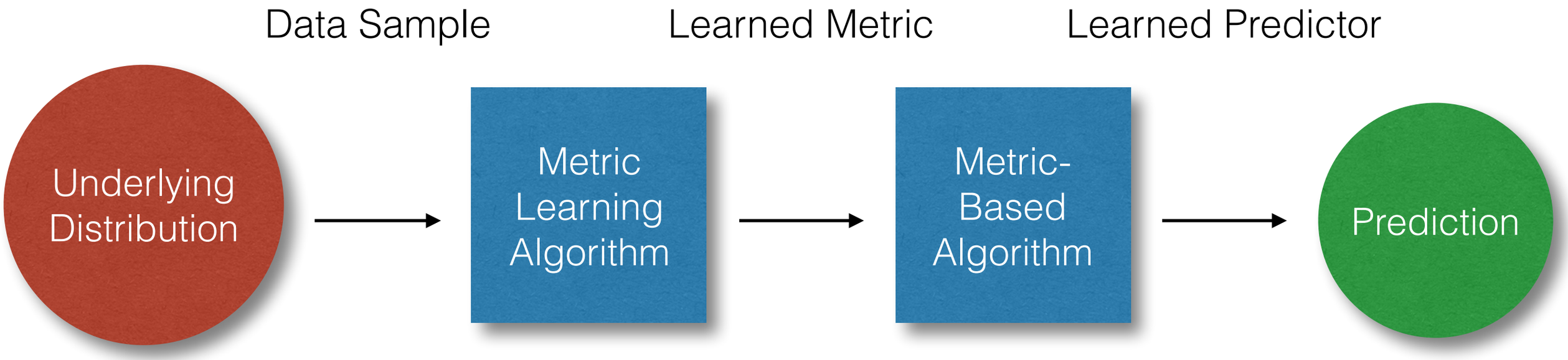
(b)

Swiss Roll Data

The Problem of Unsupervised Distance Metric Learning

A Demo

The Supervised Approach



Learning Pipeline

How to apply the learned metric?

Metric vs. Transformation

Let

$$d_M(x_i, x_j) = (x_i - x_j)^T M (x_i - x_j)$$

A is positive semi-definite (PSD), $M = A^T A$

$$\begin{aligned} d_M(x_i, x_j) &= (x_i - x_j)^T M (x_i - x_j) \\ &= (x_i - x_j)^T A^T A (x_i - x_j) \\ &= (Ax_i - Ax_j)^T (Ax_i - Ax_j) \end{aligned}$$

Metric vs. Transformation

Using a new metric in the original space is equivalent to first applying linear transformation

$$Y = AX$$

then using Euclidean distance in new space of Y.

Transformation Types

- Identity Matrix ⇔ No Transformation
- Diagonal Matrix ⇔ Scaling Only, No Rotation
  - $A = \mathrm{diag}(1/\sigma_1, 1/\sigma_2, 1/\sigma_d)$
  - $0 - 1$
- Orthogonal Matrix ⇔ Rotation Only, No Scaling
- Full Square Matrix ⇔ Scaling + Rotation
- Non-square Full Matrix ⇔ Scaling + Rotation + Dim Reduction

Dimensionality Reduction

Take eigenvector decomposition $M_{d\times d} = V\Lambda V^T$ and

$$A = (\Lambda^{1/2}V^T)^T_{d\times r}(\Lambda^{1/2}V^T)_{r\times d} \quad (r < d)$$

then

$$Y_{r\times n} = A_{r\times d}X_{d\times n}$$

Supervised Approach

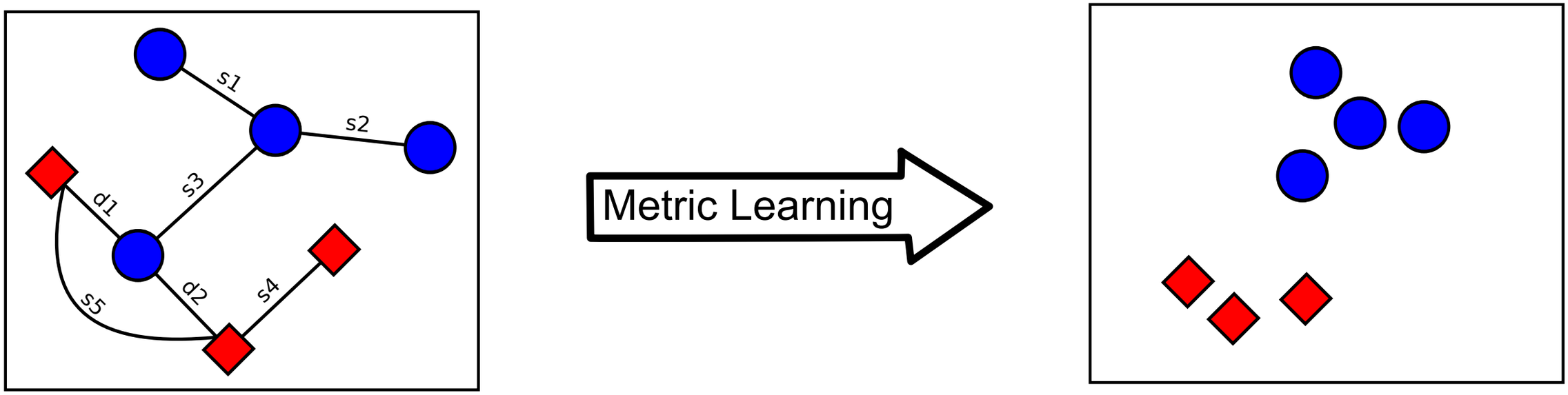
- Early Approach
- Nearest-Neighbor Driven Approach
- Information Theoretic Driven Approach

Early Approaches

Introducing Side-Information

Must-Link / Cannot Link Constrains:

$$S = \{(x_i, x_j) : x_i \text{ and } x_j \text{ should be similar}\}$$
$$D = \{(x_i, x_j) : x_i \text{ and } x_j \text{ should be dissimilar}\}$$



Intuition Behind Early Approach

Learning Form

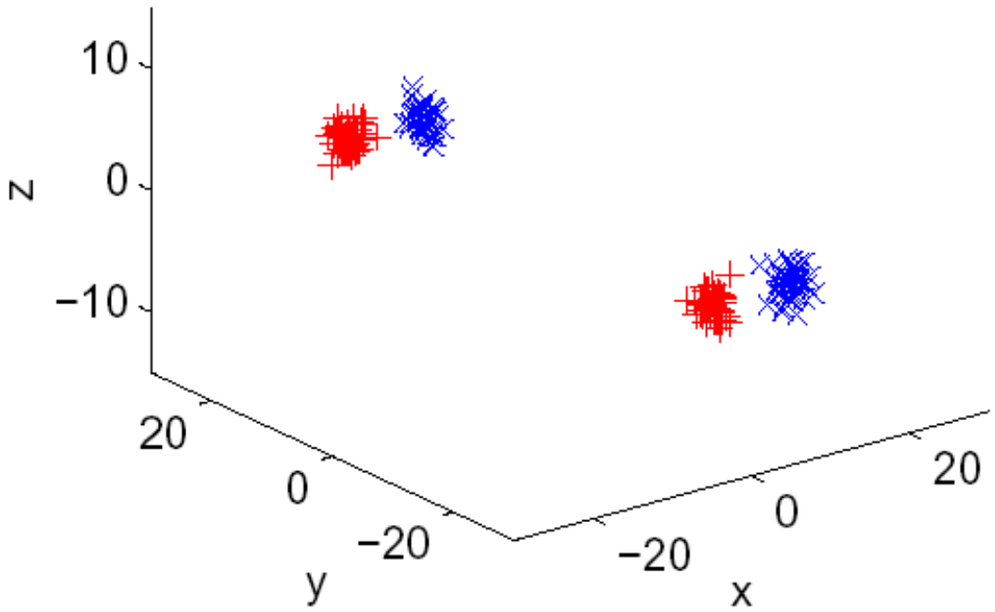
General Form:

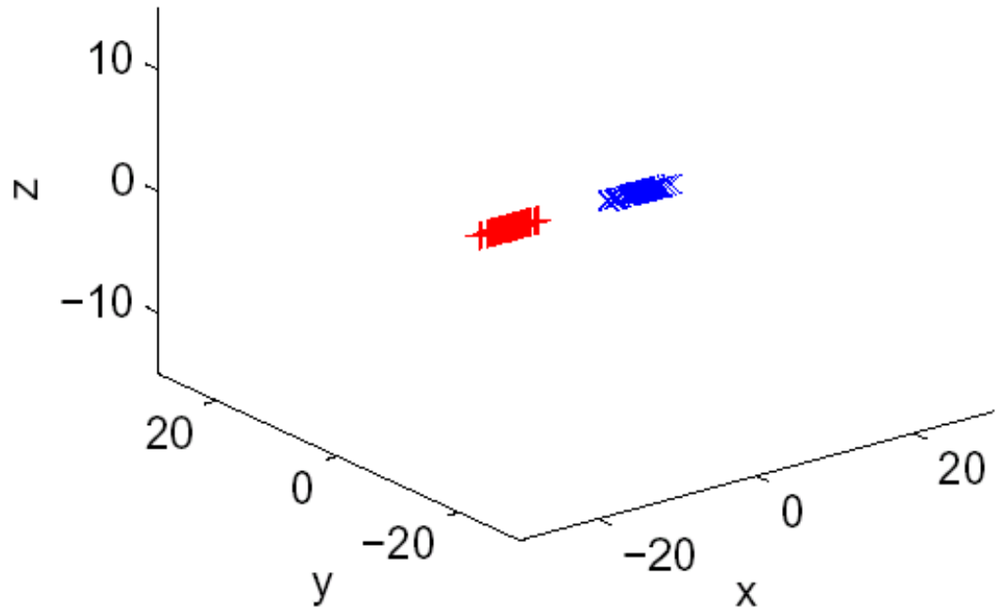
$$\min_{M} \quad \ell(M, S, D) \; + \; \lambda R(M)$$

GDMLCP (Xing et al., 2002)

$$\min_{\mathbf{M}} \sum_{(\mathbf{x}_i,\mathbf{x}_j)\in S} ||\mathbf{x}_i - \mathbf{x}_j||^2_{\mathbf{M}}$$

$$\text{s.t.} \quad \sum_{(\mathbf{x}_i,\mathbf{x}_j)\in D} ||\mathbf{x}_i - \mathbf{x}_j||_{\mathbf{M}} \geq 1, \quad \mathbf{M} \succcurlyeq 0.$$





Original Data and Projected Data

GDMLCP

- Convex Optimization
- Very Slow, Iterative, Not Parallelizable
- Global Optimum, Intractable for Multimodal Data

J. Fan et al., (Nov 2013)

QUADRO: A Supervised Dimension Reduction Method via Rayleigh Quotient Optimization

arxiv.1311.5542

Exactly the same idea,
but for high-dim classification

Nearest Neighbor Driven Approaches

NCA (Hinton et al., 2004)

Minimize the expected LOO error of a stochastic NN classifier in the projection space introduced by *M*.

NCA

The probability x_i is a neighbor of x_j:

$$p_{ij} = \frac{\exp(-||x_i - x_j||_M^2)}{\sum_{l \neq i} \exp(-||x_i - x_l||_M^2)}, \quad p_{ii} = 0.$$

The probability x_i is correctly classified:

$$p_i = \sum_{j:y_j = y_i} p_{ij}$$

then learn the distance metric by solving:

$$\max_M \sum_i p_i$$

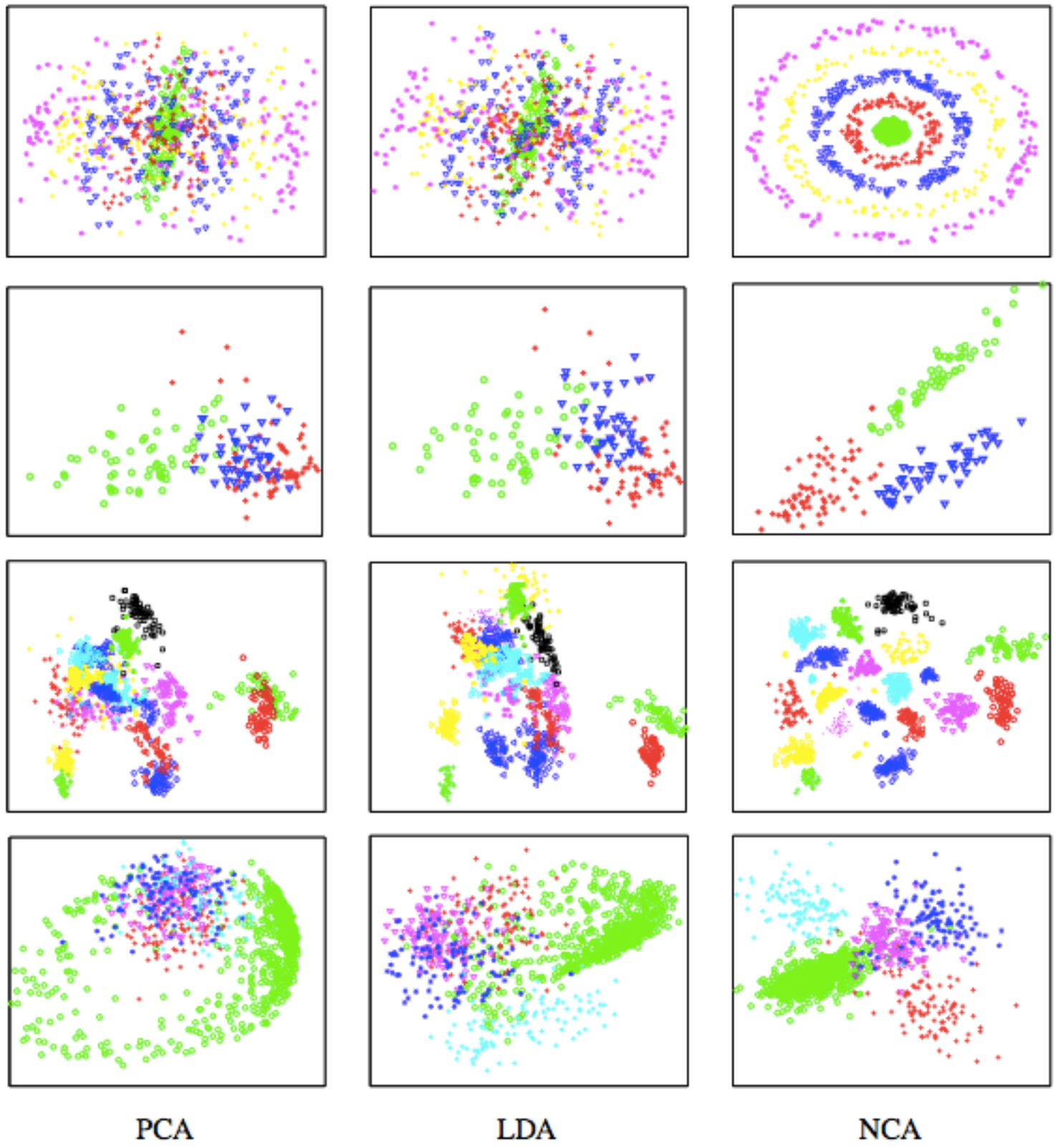


Figure 2: Dataset visualization results of PCA, LDA and NCA applied to (from top) the "concentric rings", "wine", "faces" and "digits" datasets. The data are reduced from their original dimensionalities (D=3,D=13,D=560,D=256 respectively) to the d=2 dimensions

NCA Summary

- Non-Convex Optimization
- Local Maxima

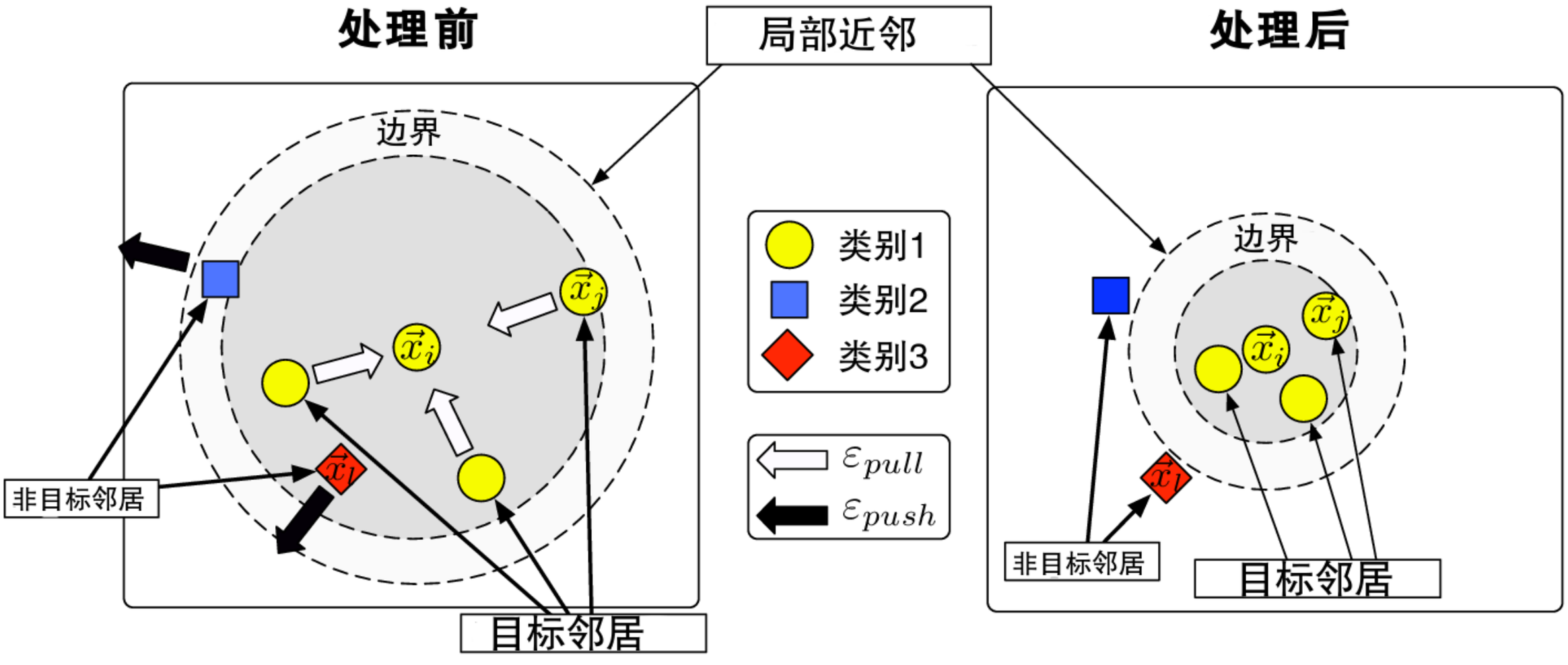
LMNN (Weinberger, 2005)

- Most widely used, many variants

- Local constrains inspired by SVM:

The k-nearest Euclidean neighbors (“target neighbors”) of any training samples should belong to the correct class while keeping away samples of other classes.

LMNN


处理前
局部近邻
处理后
边界
$\vec{x}_j$
$\vec{x}_i$
$\vec{x}_l$
非目标邻居
目标邻居
类别1
类别2
类别3
ε_{pull}
ε_{push}
边界
$\vec{x}_j$
$\vec{x}_i$
$\vec{x}_l$
非目标邻居
目标邻居

LMNN

$$S = \{(x_i, x_j) : y_i = y_j \text{ and } x_j \text{ belongs to the k-neighborhood of } x_i\}$$

$$R = \{(x_i, x_j, x_k) : (x_i, x_j) \in S,\ y_i \neq y_k\}$$

The metric is learned by the convex optimization:

$$\min_M \sum_{(x_i, x_j) \in S} (1-\mu)\, d_M^2(x_i, x_j) \ + \ \mu \sum_{i,j,k} \xi_{i,j,k}$$

$$\text{s.t.} \quad d_M^2(xi, x_k) - d_M^2(x_i, x_j) \geq 1 - \xi_{ijk} \qquad \forall (x_i, x_j, x_k) \in R$$

LMNN Summary

- Convex Optimization
- Special Purpose Solver by Subgradient Descent
- Parallelizable, Scales to Billions of Constrains
- Constrained by Euclidean NN Selection

Information Theoretic Driven Approaches

RCA (Bar-Hillel, 2003)

"Chunklets"

$$\hat{C} = \frac{1}{n} \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{ji} - \hat{m}_j)(x_{ji} - \hat{m}_j)^T$$

The inverse of $\hat{C}$ is a Mahalanobis distance metric.

RCA

- It is a optimal solution to a mutual information measure.

- It is a optimal solution to the optimization problem consisting in minimizing the within-class distances.

RCA (cont'd)

- DCA (Discriminant Component Analysis) for incorporating cannot link constrains

- Kernel RCA

ITML (Davis, 2007)

Define Bregman Divergence on a PSD matrix:

$$D_{ld}(M, M_0) = \mathrm{tr}(MM_0^{-1}) - \log\det(MM_0^{-1}) - d$$

M_0 is often set to I thus the regularization aims at keeping the learned distance close to Euclidean distance.

ITML

$$\min_{M} D_{ld}(M, M_0) \;+\; \gamma \sum_{i,j} \xi_{i,j}$$

$$\text{s.t.} \quad d^2_M(x_i, x_j) \leq \mu + \xi_{i,j} \quad \forall (x_i, x_j) \in S$$

$$d^2_M(x_i, x_j) \geq \nu - \xi_{i,j} \quad \forall (x_i, x_j) \in D$$

ITML Summary

- The information-theoretic interpretation behind minimizing D_{ld} is equivalent to minimizing the KL divergence between two multivariate Gaussian distributions parameterized by M and M_0.

ITML Summary

- Efficient Optimization Algorithm & Global Optimum
- Hand-picked M_0 may have a unknown influence

Applications

Music Classification

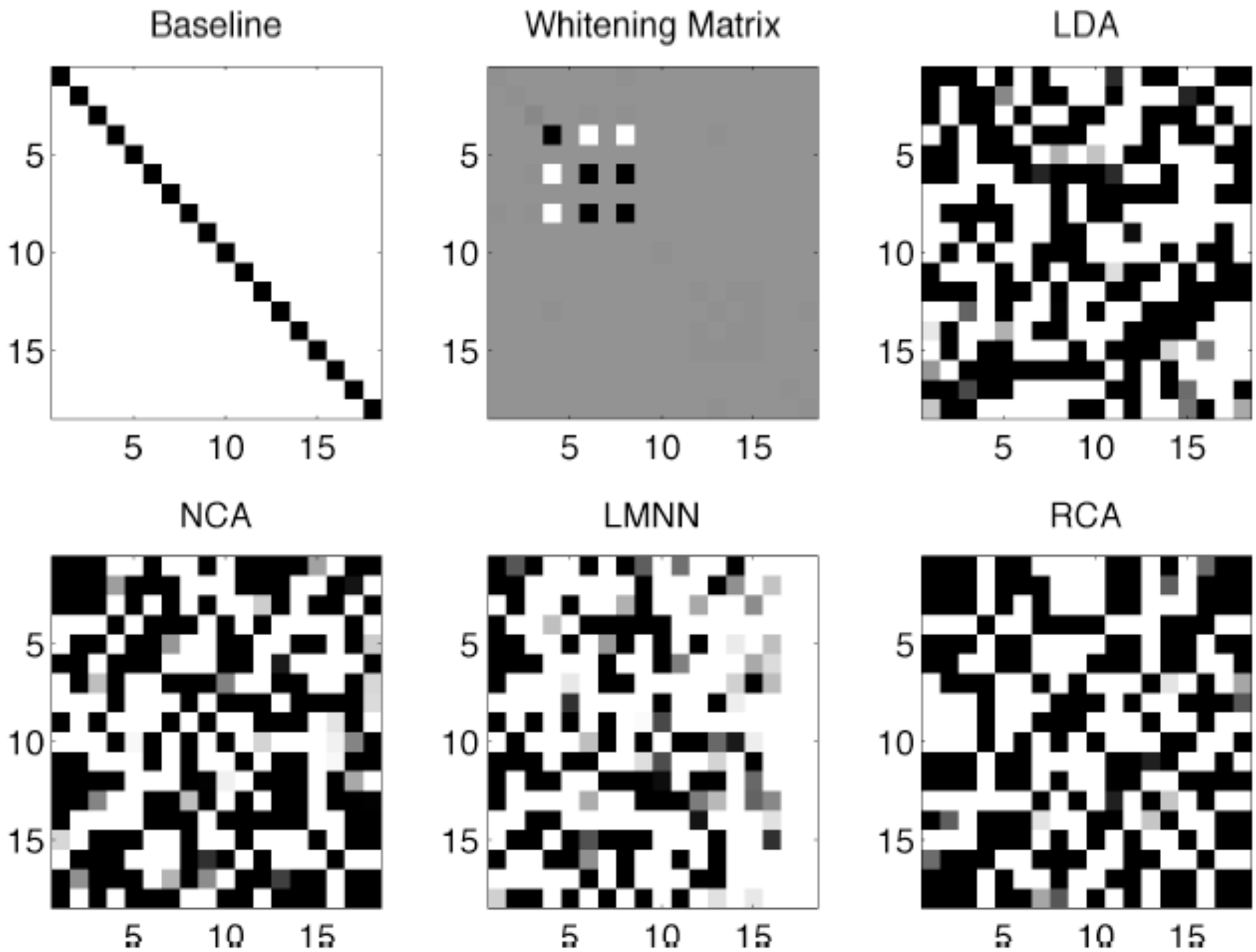
Task Briefings

- MP3s downloaded from a set of music blogs
  - Thousands of songs
  - 319 blogs, 164 artists, 74 albums

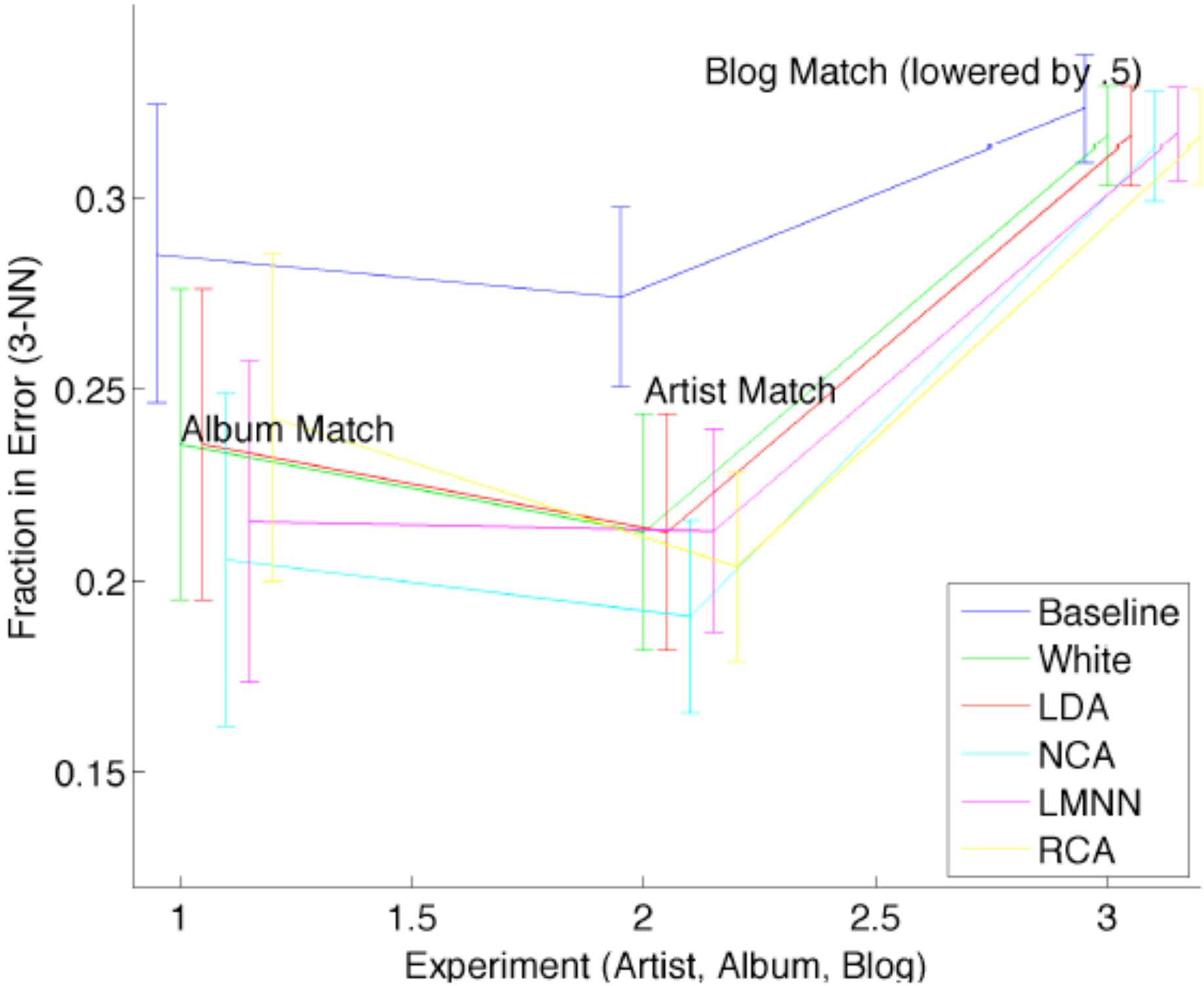
Slaney, M., Weinberger, K. Q., & White, W. (2008). Learning a Metric for Music Similarity.

Task Briefings

- Extract 18 features for each song:
    - Songs broken up into segments (80ms to a few seconds)
    - Mean segment duration
    - Track tempo estimate
    - Regularity of the beat
    - Estimation of the time signature overall loudness estimate of the track
    - ……
- Training done via labels based on blog, artist, and album (separately)



Learned Distance Matrix

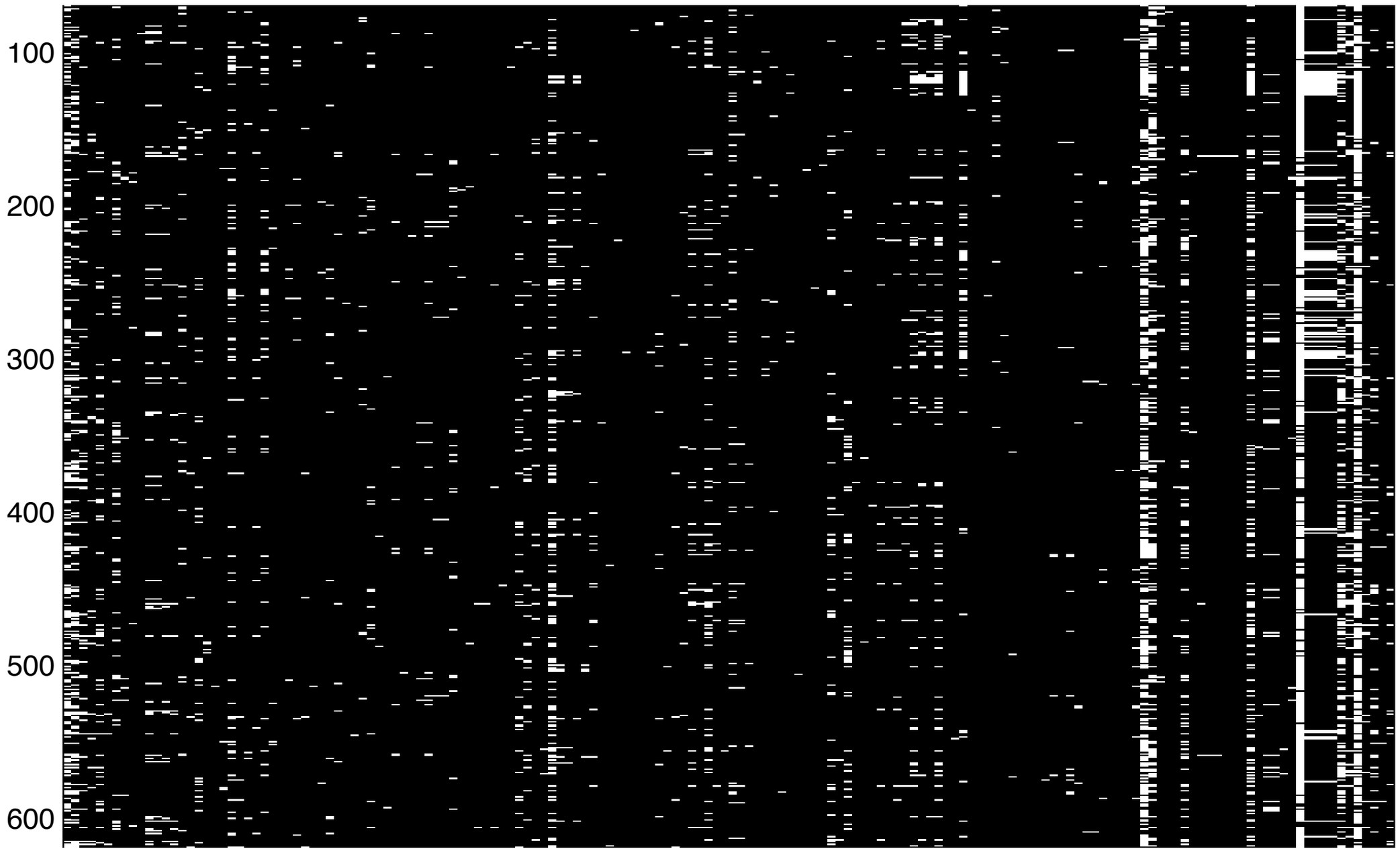


Performance Comparison

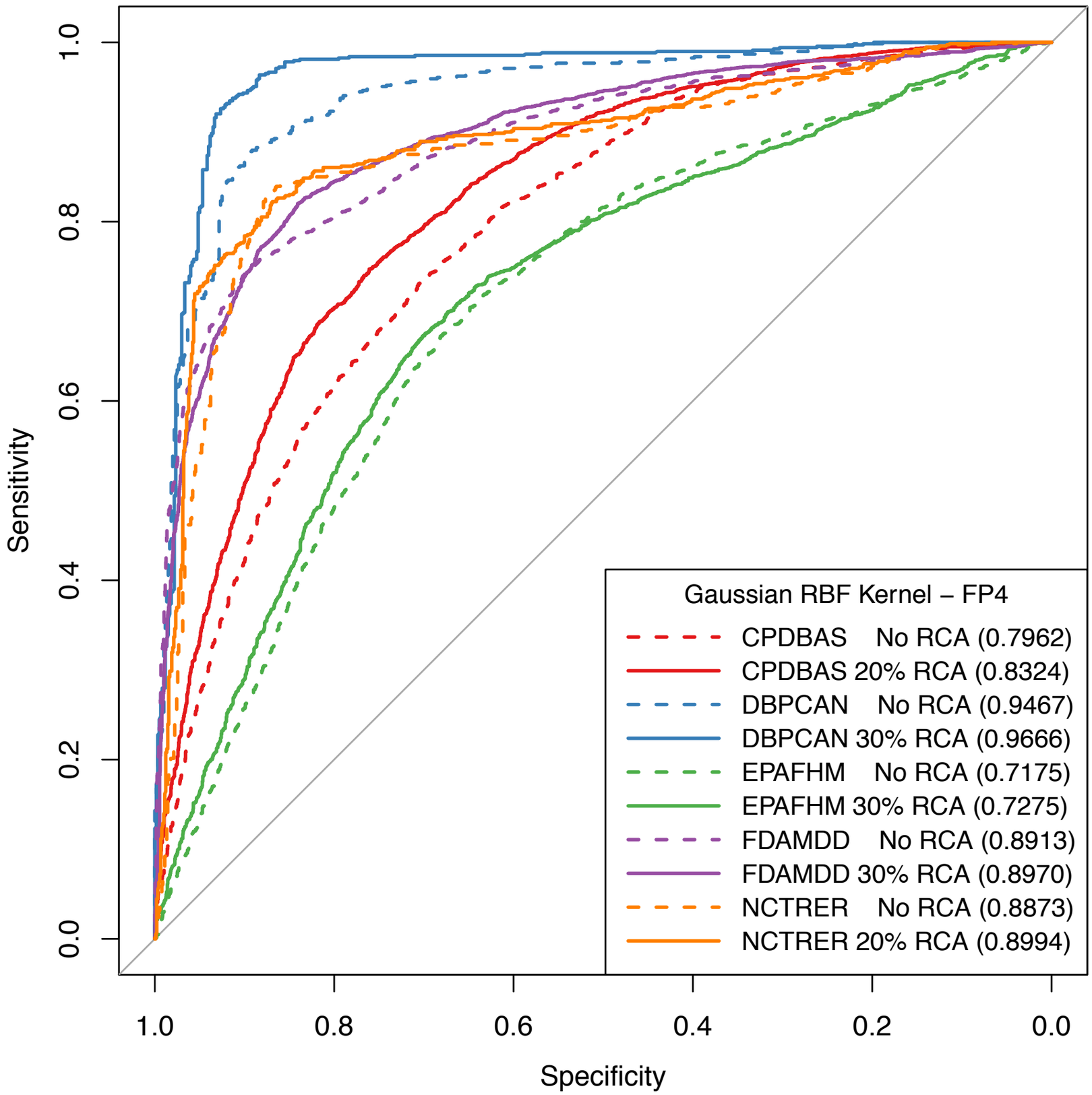
In-Silico Toxicity Prediction

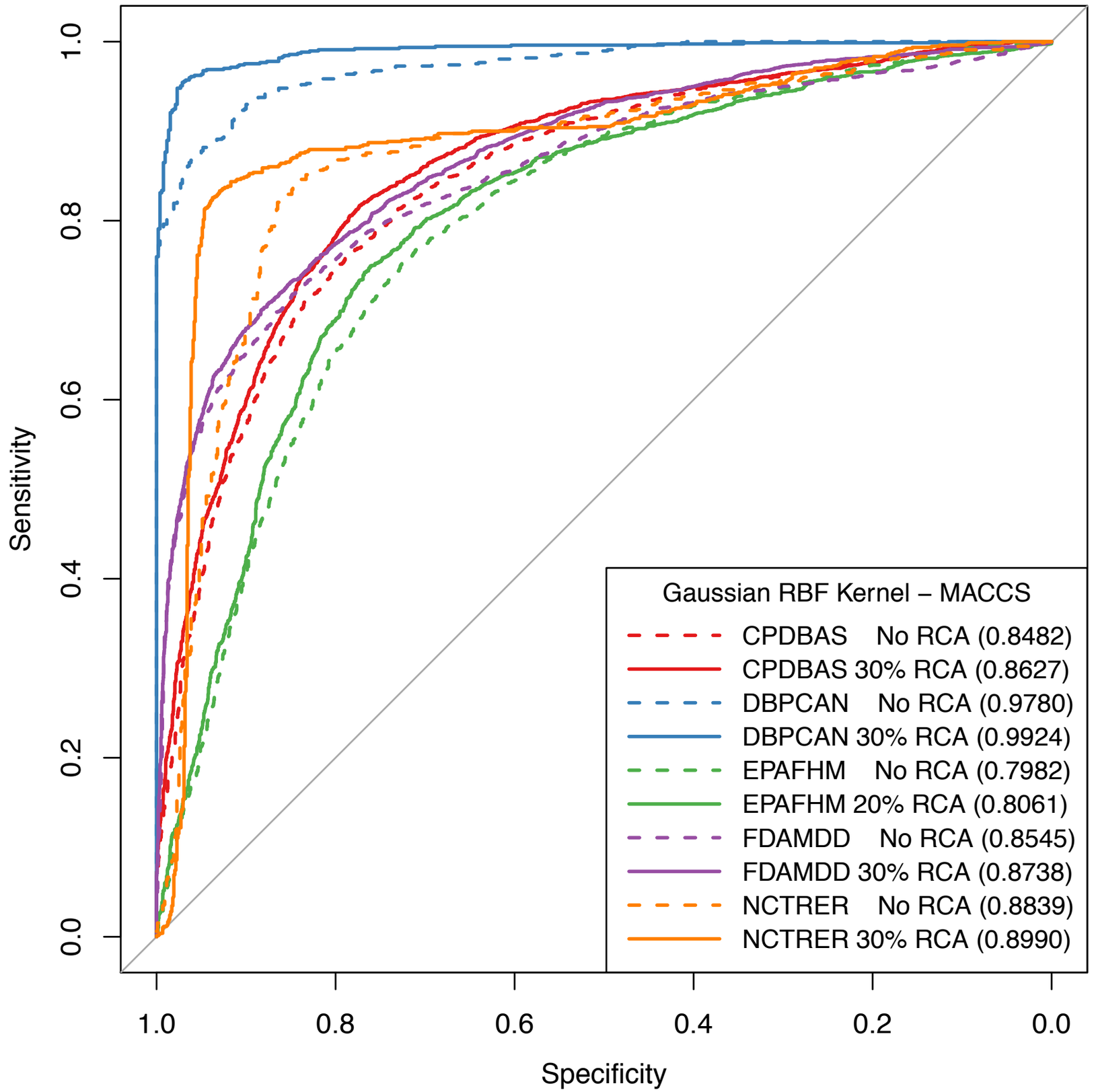
Task Briefings

- Five toxicity datasets from FDA, EPA, etc.
- Toxic / Non-Toxic, 100 / 100 samples
- FP4 / MACCS 0-1 valued molecular fingerprints



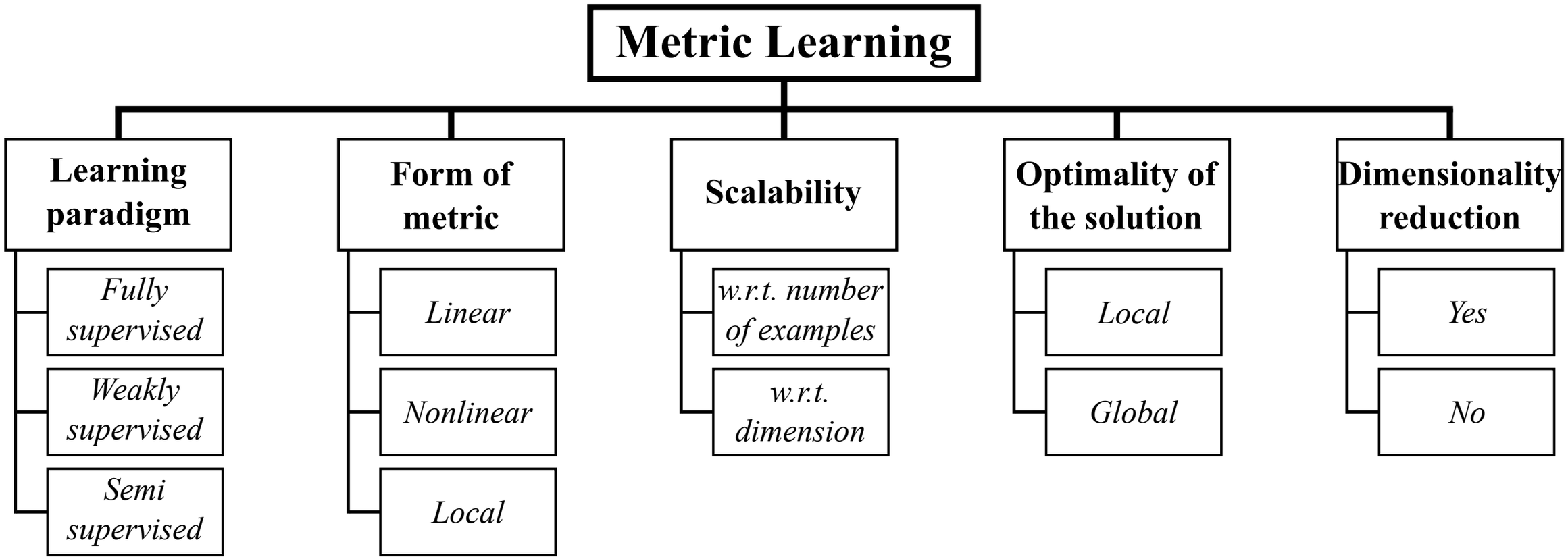
A Glimpse of Data


Sensitivity
1.0
0.8
0.6
0.4
0.2
0.0
Specificity
1.0
0.8
0.6
0.4
0.2
0.0
Gaussian RBF Kernel – FP4
CPDBAS No RCA (0.7962)
CPDBAS 20% RCA (0.8324)
DBPCAN No RCA (0.9467)
DBPCAN 30% RCA (0.9666)
EPAFHM No RCA (0.7175)
EPAFHM 30% RCA (0.7275)
FDAMDD No RCA (0.8913)
FDAMDD 30% RCA (0.8970)
NCTRER No RCA (0.8873)
NCTRER 20% RCA (0.8994)


Sensitivity
1.0
0.8
0.6
0.4
0.2
0.0
Specificity
1.0
0.8
0.6
0.4
0.2
0.0
Gaussian RBF Kernel – MACCS
CPDBAS No RCA (0.8482)
CPDBAS 30% RCA (0.8627)
DBPCAN No RCA (0.9780)
DBPCAN 30% RCA (0.9924)
EPAFHM No RCA (0.7982)
EPAFHM 20% RCA (0.8061)
FDAMDD No RCA (0.8545)
FDAMDD 30% RCA (0.8738)
NCTRER No RCA (0.8839)
NCTRER 30% RCA (0.8990)

Model Summary

- Fully- / Weakly- / Semi-Supervised
- Metric Form: Linear (Kernelization) / Non-Linear / Local
- Optimum: Global / Local
- Dim Reduction / Non-Dim Reduction
- Regularization / No Regularization



5 Key Properties of Supervised Distance Metric Learning Algorithms

Name	Year	Source Code	Supervision	Form of Metric	Scalability w.r.t. n	Scalability w.r.t. d	Optimum	Dimension Reduction	Regularizer	Additional Information
MMC	2002	Yes	Weak	Linear	★☆☆	☆☆☆	Global	No	None	—
S&J	2003	No	Weak	Linear	★☆☆	★★☆	Global	No	Frobenius	—
NCA	2004	Yes	Full	Linear	★☆☆	★★☆	Local	Yes	None	For k-NN
MCML	2005	Yes	Full	Linear	★☆☆	☆☆☆	Global	No	None	For k-NN
LMNN	2005	Yes	Full	Linear	★★☆	★☆☆	Global	No	None	For k-NN
RCA	2003	Yes	Weak	Linear	★★☆	★★☆	Global	No	None	—
ITML	2007	Yes	Weak	Linear	★☆☆	★★☆	Global	Yes	LogDet	Online version
SDML	2009	No	Weak	Linear	★☆☆	★★☆	Global	Yes	LogDet+L_1	$n \ll d$
POLA	2004	No	Weak	Linear	★★★	★☆☆	Global	No	None	Online
LEGO	2008	No	Weak	Linear	★★★	★★☆	Global	Yes	LogDet	Online
RDML	2009	No	Weak	Linear	★★★	★★☆	Global	No	Frobenius	Online
MDML	2012	No	Weak	Linear	★★★	★☆☆	Global	Yes	Nuclear norm	Online
mt-LMNN	2010	Yes	Full	Linear	★★☆	☆☆☆	Global	No	Frobenius	Multi-task
MLCS	2011	No	Weak	Linear	★☆☆	★★☆	Local	Yes	N/A	Multi-task
GPML	2012	No	Weak	Linear	★☆☆	★★☆	Global	Yes	von Neumann	Multi-task
TML	2010	Yes	Weak	Linear	★★☆	★★☆	Global	No	Frobenius	Transfer learning
LPML	2006	No	Weak	Linear	★★☆	★★☆	Global	Yes	L_1	—
SML	2009	No	Weak	Linear	★☆☆	☆☆☆	Global	Yes	$L_{2,1}$	—
BoostMetric	2009	Yes	Weak	Linear	★☆☆	★★☆	Global	Yes	None	—
DML-p	2012	No	Weak	Linear	★☆☆	★☆☆	Global	No	None	—
RML	2010	No	Weak	Linear	★★☆	☆☆☆	Global	No	Frobenius	Noisy constraints
MLR	2010	Yes	Full	Linear	★★☆	☆☆☆	Global	Yes	Nuclear norm	For ranking
SiLA	2008	No	Full	Linear	★★☆	★★☆	N/A	No	None	Online
gCosLA	2009	No	Weak	Linear	★★★	☆☆☆	Global	No	None	Online
OASIS	2009	Yes	Weak	Linear	★★★	★★☆	Global	No	Frobenius	Online
SLLC	2012	No	Full	Linear	★★☆	★★☆	Global	No	Frobenius	For linear classif.
RSL	2013	No	Full	Linear	★☆☆	★★☆	Local	No	Frobenius	Rectangular matrix
LSMD	2005	No	Weak	Nonlinear	★☆☆	★★☆	Local	Yes	None	—
SVML	2012	No	Full	Nonlinear	★☆☆	★★☆	Local	Yes	Frobenius	For SVM
GB-LMNN	2012	No	Full	Nonlinear	★★☆	★★☆	Local	Yes	None	—
M^2-LMNN	2008	Yes	Full	Local	★★☆	★☆☆	Global	No	None	—
GLML	2010	No	Full	Local	★★★	★★☆	Global	No	Diagonal	Generative
Bk-means	2009	No	Weak	Local	★☆☆	★★★	Global	No	RKHS norm	Bregman dist.
PLML	2012	Yes	Weak	Local	★★☆	☆☆☆	Global	No	Manifold+Frob	—
RFD	2012	Yes	Weak	Local	★★☆	★★★	N/A	No	None	Random forests
χ^2-LMNN	2012	No	Full	Nonlinear	★★☆	★★☆	Local	Yes	None	Histogram data
GML	2011	No	Weak	Linear	★☆☆	★★☆	Local	No	None	Histogram data
EMDL	2012	No	Weak	Linear	★☆☆	★★☆	Local	No	Frobenius	Histogram data
LRML	2008	Yes	Semi	Linear	★☆☆	☆☆☆	Global	No	Laplacian	—
M-DML	2009	No	Semi	Linear	★☆☆	☆☆☆	Local	No	Laplacian	Auxiliary metrics
SERAPH	2012	Yes	Semi	Linear	★☆☆	☆☆☆	Local	Yes	Trace+entropy	Probabilistic
CDML	2011	No	Semi	N/A	N/A	N/A	N/A	N/A	N/A	Domain adaptation
DAML	2011	No	Semi	Nonlinear	★☆☆	☆☆☆	Global	No	MMD	Domain adaptation

Future Problems

- Scalability with both *p* and *n* — $O(p^2)$
- High Dim and Sparse Metric Learning (J-L Lemma)
- Learning Richer Metrics — Similarity Degree?
- Metric Learning for Structured Data (Char/Tree/Graph)
- CV / NLP / Bio- / Chem- / -Informatics / -Metrics

References

- (Survey) A. Bellet. A. Habrard. M. Sebban. (2013). A Survey on Metric Learning for Feature Vectors and Structured Data.

- (Survey) Brian Kulis. (2012). Metric Learning: A Survey.

- (Survey) Yang Liu. (2007). Distance Metric Learning: A Comprehensive Survey.

- (GDMLCP) Eric P. Xing, Andrew Y. Ng, Michael I. Jordan, Stuart Russell. (2003). Distance Metric Learning with Application to Clustering with Side-information. NIPS'15.

- (NCA) J. Goldberger, S. Roweis, G. Hinton, R. Salakhutdinov. (2004). Neighborhood Components Analysis. NIPS'17.

- (LMNN) K. Q. Weinberger, J. Blitzer, L. K. Saul (2006). Distance Metric Learning for Large Margin Nearest Neighbor Classification. NIPS'18.

- (RCA) Aharon Bar-Hillel, Tomer Hertz, Noam Shental, Daphna Weinshall. (2003). Learning Distance Functions Using Equivalence Relations. ICML'03.

- (ITML) Jason V. Davis, Brian Kulis, Prateek Jain, Suvrit Sra, Inderjit S. Dhillon. (2007). Information-theoretic Metric Learning. ICML'07.

Q & A