Structural pattern recognition in dissimilarity space

Robert P.W. Duin, Delft University of Technology

Pattern Recognition Lab
Delft University of Technology, The Netherlands
http://rduin.nl

TUDelft

Structural pattern recognition

- Shapes
- Sequences
- Graphs

Objects → Feature representation → Classifier

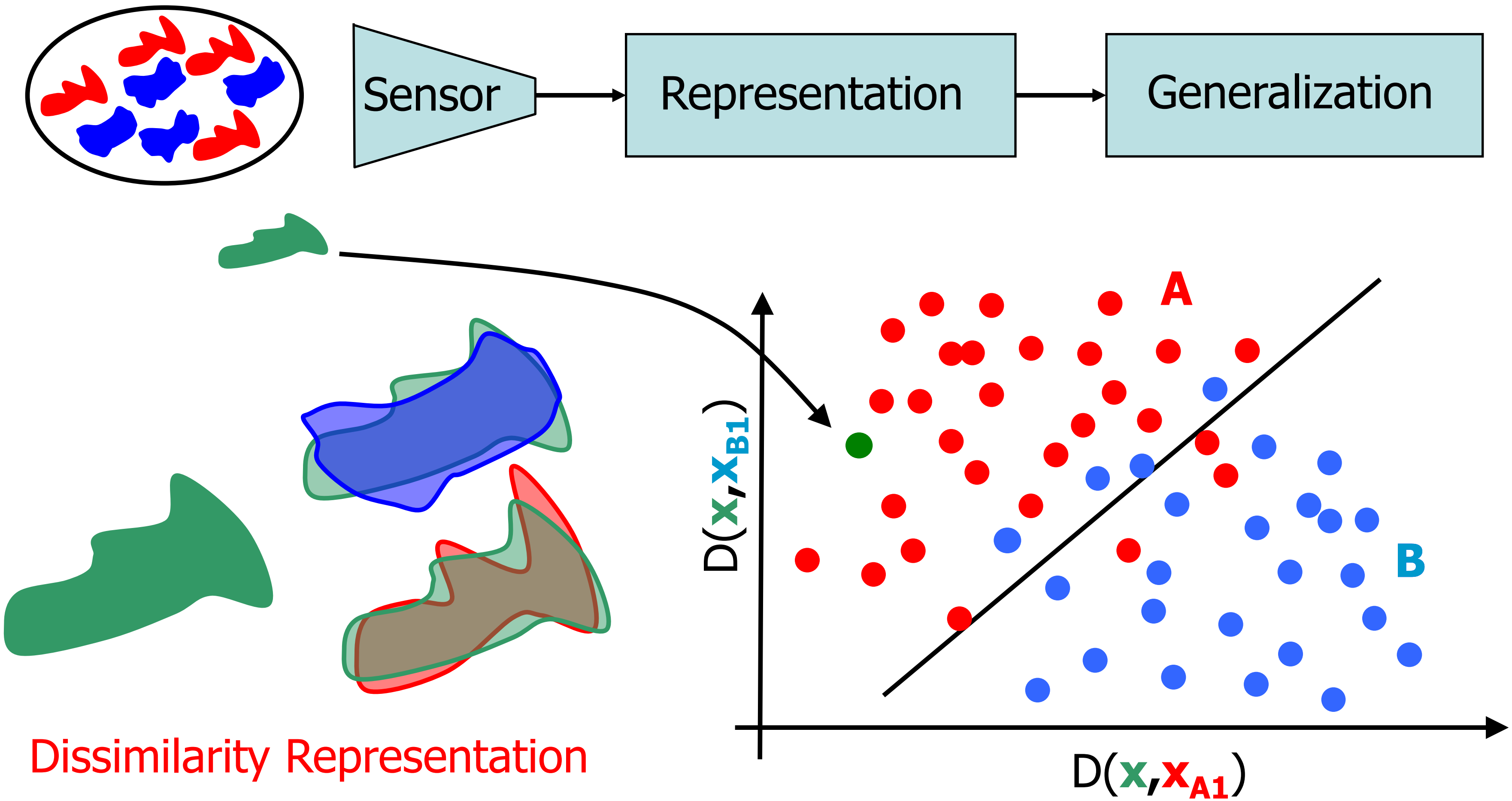
Objects → Similarity / Dissimilarity → Nearest Neighbor Rule

Here:

Objects → Dissimilarities → Dissimilarity Space → Classifier

TU Delft

Dissimilarities → True Representation



Dissimilarity Representation

TUDelft

Dissimilarity Measures

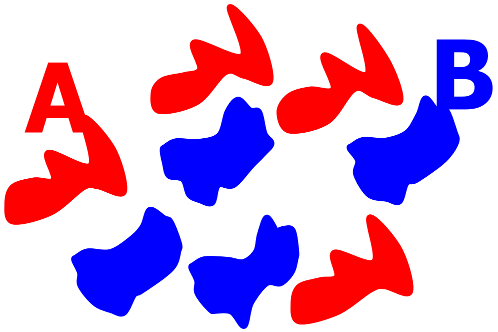
- Shapes : Geometrics, Morphing, Editing
- Sequences: DTW, HMM, Editing
- Graphs: Graph distances based on nodes / edges /attributes

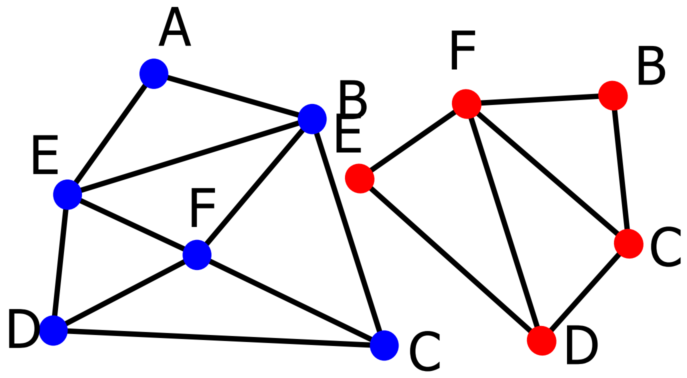
Questions

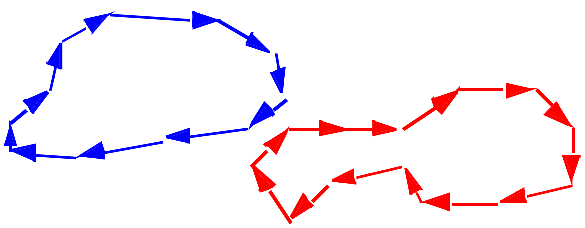
How to improve a given dissimilarity representation?

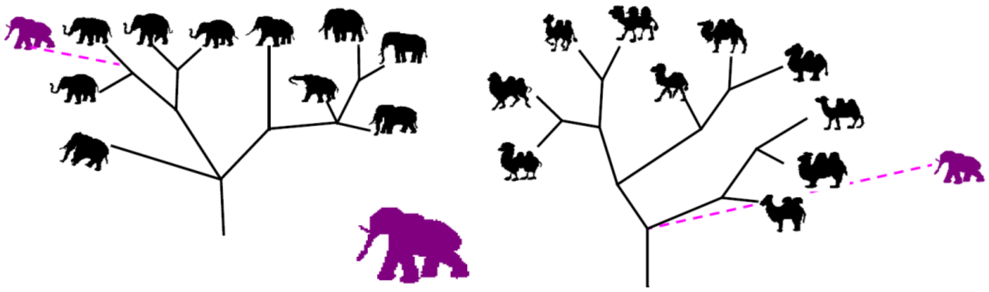
Should it be Euclidean? (~Mercer kernels)

Structural Representation



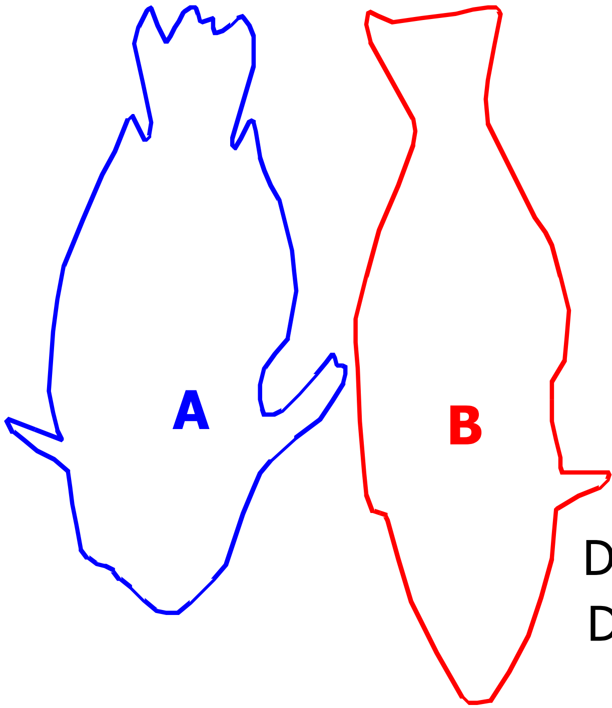




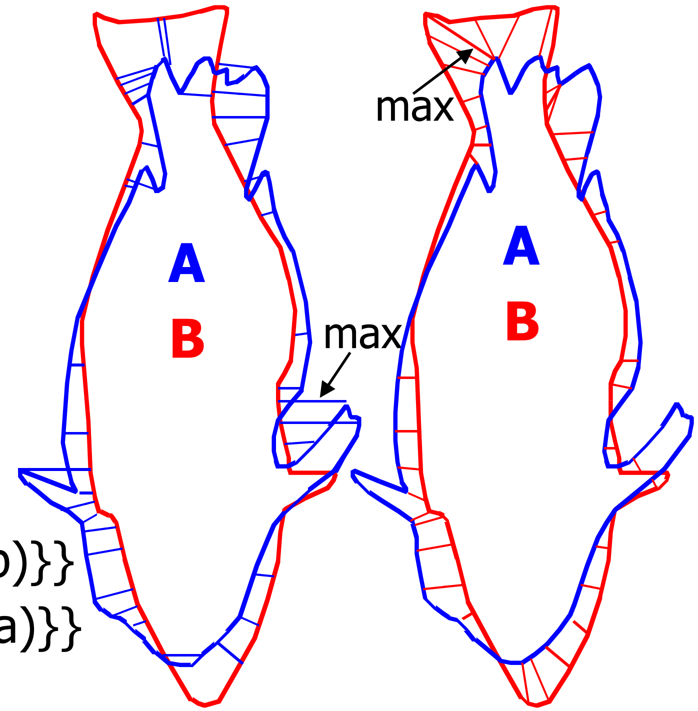


How to generalize? Distances!

Examples Dissimilarity Measures



Dist(**A**,**B**):
a ∈ **A**, points of **A**
b ∈ **B**, points of **B**
d(a,b): Euclidean distance



D(**A**,**B**) = max_a{min_b{d(a,b)}}
D(**B**,**A**) = max_b{min_a{d(b,a)}}

Hausdorff Distance (metric):
DH = max{max_a{min_b{d(a,b)}} , max_b{min_a{d(b,a)}}}

D(**A**,**B**) ≠ D(**B**,**A**)

Modified Hausdorff Distance (non-metric):
DM = max{mean_a{min_b{d(a,b)}},mean_b{min_a{d(b,a)}}}

Dubuisoon & Jain, Modified Hausdorff distance for object matching, ICPR12, 2004,, voll 1, 566-568.

Dissimilarities – Possible Assumptions

Metric

1. Positivity: $d_{ij} \geq 0$
2. Reflexivity: $d_{ii} = 0$
3. Definiteness: $d_{ij} = 0$ iff objects i and j are identical
4. Symmetry: $d_{ij} = d_{ji}$
5. Triangle inequality: $d_{ij} < d_{ik} + d_{kj}$
6. Compactness: if the objects i and j are very similar then $d_{ij} < \delta$.
7. True representation: if $d_{ij} < \delta$ then the objects i and j are very similar.
8. Continuity of d.

TU Delft

Class separability

The **identity property**:

- Definiteness: $d_{ij} = 0$ iff objects i and j are identical

causes no-overlapping classes if objects are uniquely labeled.

There might be entirely different dissimilarity measures that have this property. Combining helps?

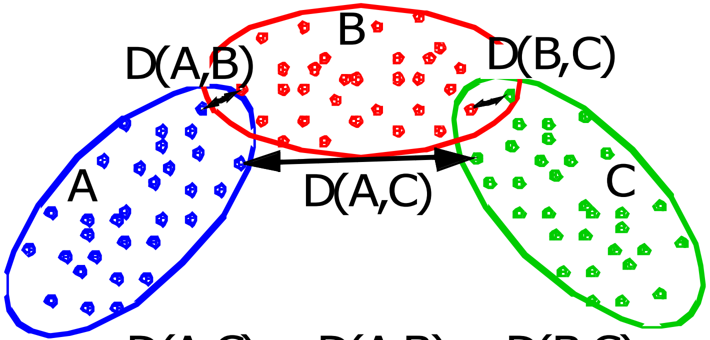
The identity property is sometimes not fulfilled

Distance(Table,Book) = 0

Distance(Table,Cup) = 0

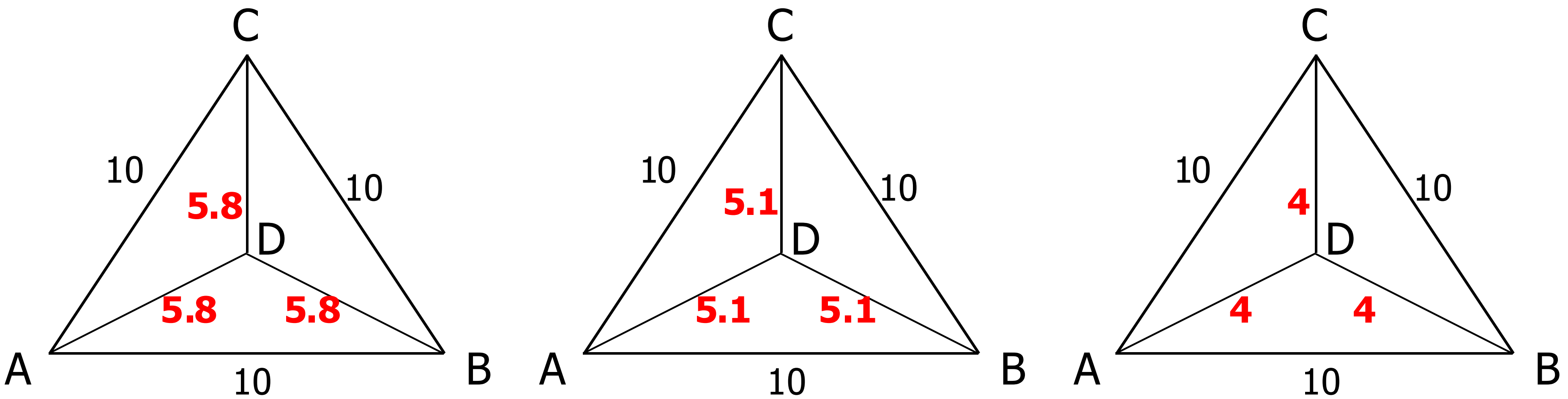
Distance(Book,Cup) = 1

Single-linkage clustering



D(A,C) > D(A,B) + D(B,C)

Euclidean - Non Euclidean - Non Metric



Euclidean metric

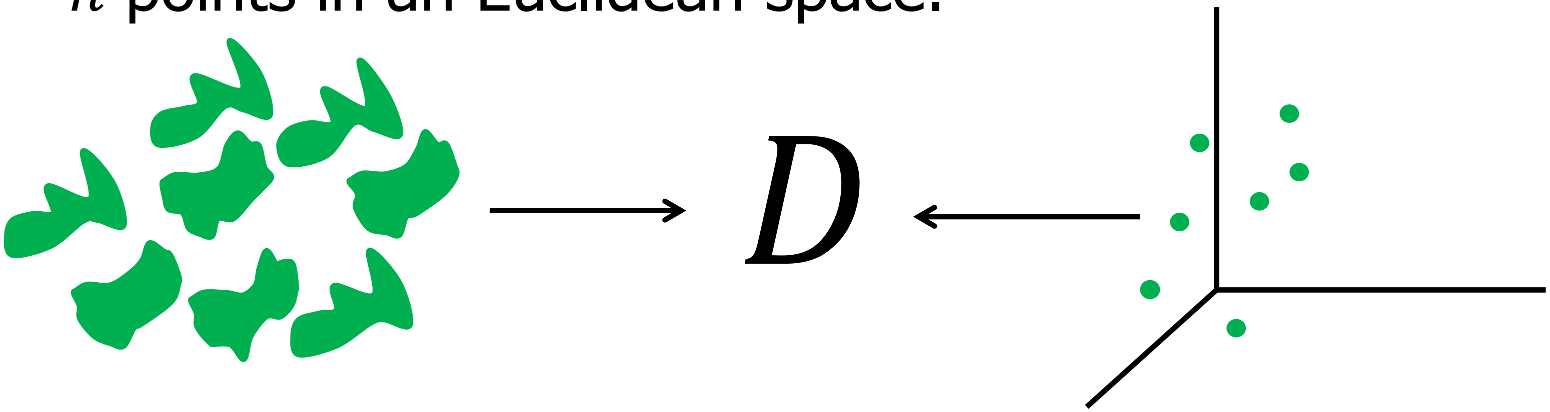
non-Euclidean metric

non-Euclidean non-metric

What is an Euclidean dissimilarity matrix?

Definition

An $n \times n$ dissimilarity matrix between n objects is Euclidean if it can arise as the distances between n points in an Euclidean space.



Note: An Euclidean dissimilarity matrix is **square** and has a **zero diagonal**, but this **insufficient** to be Euclidean.

TU Delft

Euclidean dissimilarities

Theorem I:

Let D_1^2 and D_2^2 be squared Euclidean dissimilarity matrices, then:

$$D^2 = \alpha D_1^2 + \beta D_2^2 \quad (\alpha, \beta \geq 0)$$

is a squared Euclidean dissimilarity matrix as well.

TU Delft

Non-Euclidean Dissimilarities

<u>Theorem II:</u>

Let D^2 be a squared non-Euclidean distance matrix, symmetric and with zero diagonal, then:

$$D^2 = D_p^2 - D_n^2$$

such that

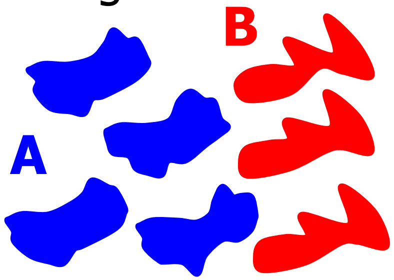
$$D_p^2 \text{ and } D_n^2$$

are both Euclidean

TU Delft

Alternatives for the Nearest Neighbor Rule

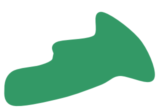
Training set



Dissimilarities d_{ij} between all training objects

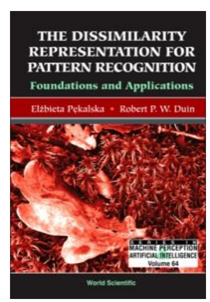
$$D_T = \begin{pmatrix} d_{11} & d_{12} & d_{13} & d_{14} & d_{15} & d_{16} & d_{17} \\ d_{21} & d_{22} & d_{23} & d_{24} & d_{25} & d_{26} & d_{27} \\ d_{31} & d_{32} & d_{33} & d_{34} & d_{35} & d_{36} & d_{37} \\ d_{41} & d_{42} & d_{43} & d_{44} & d_{45} & d_{46} & d_{47} \\ d_{51} & d_{52} & d_{53} & d_{54} & d_{55} & d_{56} & d_{57} \\ d_{61} & d_{62} & d_{63} & d_{64} & d_{65} & d_{66} & d_{67} \\ d_{71} & d_{72} & d_{73} & d_{74} & d_{75} & d_{76} & d_{77} \end{pmatrix}$$

Unlabeled object **x** to be classified



$$d_x = (d_{x1} \; d_{x2} \; d_{x3} \; d_{x4} \; d_{x5} \; d_{x6} \; d_{x7})$$

1. Dissimilarity Space
2. Embedding



Pekalska, The dissimilarity representation for PR. World Scientific, 2005.

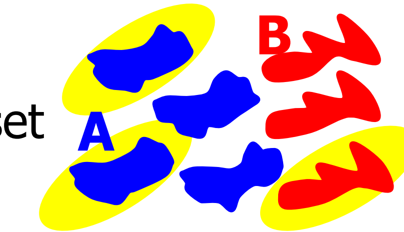
Alternative 1: Dissimilarity Space

Dissimilarities

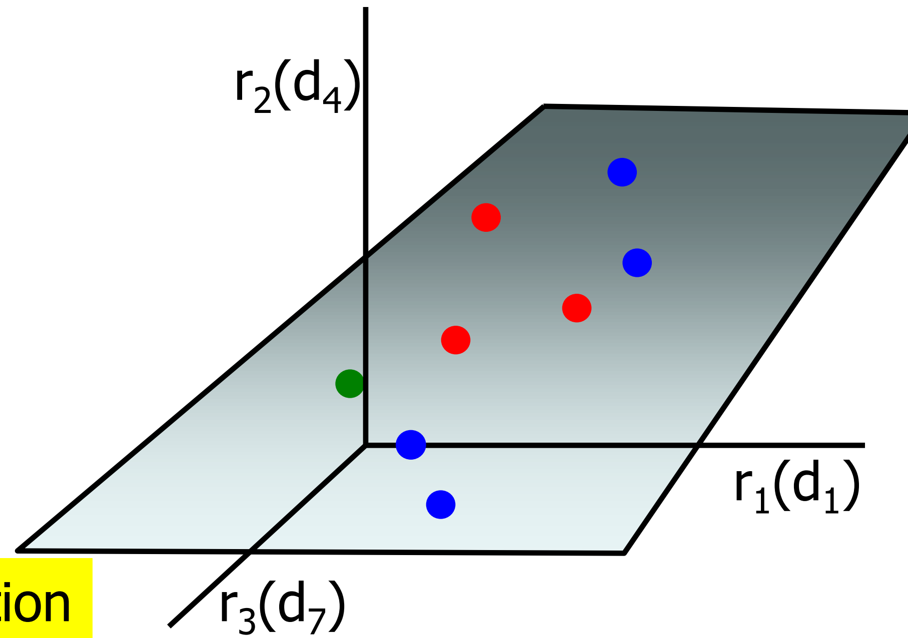
$$
\begin{array}{cccccccc}
 & r_1 & & & r_2 & & & r_3 \\
D_T = & \left(\begin{array}{c} d_{11} \\ d_{21} \\ d_{31} \\ d_{41} \\ d_{51} \\ d_{61} \\ d_{71} \end{array} \right. & \begin{array}{c} d_{12} \\ d_{22} \\ d_{32} \\ d_{42} \\ d_{52} \\ d_{62} \\ d_{72} \end{array} & \begin{array}{c} d_{13} \\ d_{23} \\ d_{33} \\ d_{43} \\ d_{53} \\ d_{63} \\ d_{73} \end{array} & \begin{array}{c} d_{14} \\ d_{24} \\ d_{34} \\ d_{44} \\ d_{54} \\ d_{64} \\ d_{74} \end{array} & \begin{array}{c} d_{15} \\ d_{25} \\ d_{35} \\ d_{45} \\ d_{55} \\ d_{65} \\ d_{75} \end{array} & \begin{array}{c} d_{16} \\ d_{26} \\ d_{36} \\ d_{46} \\ d_{56} \\ d_{66} \\ d_{76} \end{array} & \left. \begin{array}{c} d_{17} \\ d_{27} \\ d_{37} \\ d_{47} \\ d_{57} \\ d_{67} \\ d_{77} \end{array} \right)
\end{array}
$$

$$
d_x = (d_{x1} \; d_{x2} \; d_{x3} \; d_{x4} \; d_{x5} \; d_{x6} \; d_{x7})
$$

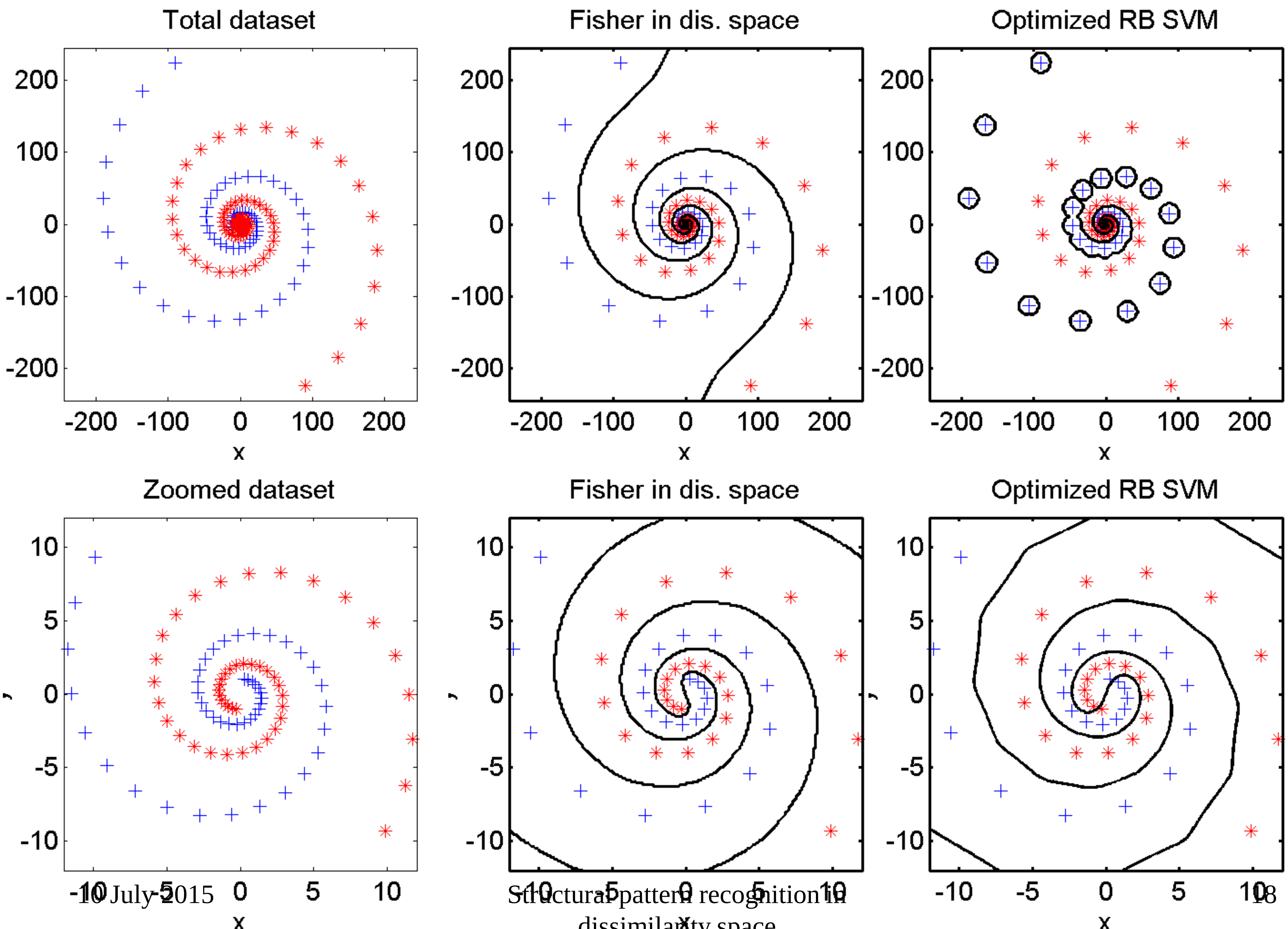
Given labeled training set



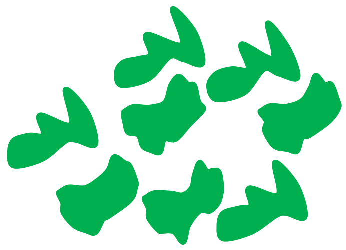
Unlabeled object to be classified



Selection of 3 objects for representation


Total dataset
Fisher in dis. space
Optimized RB SVM
Zoomed dataset
Fisher in dis. space
Optimized RB SVM
x
x
x
x
x
x

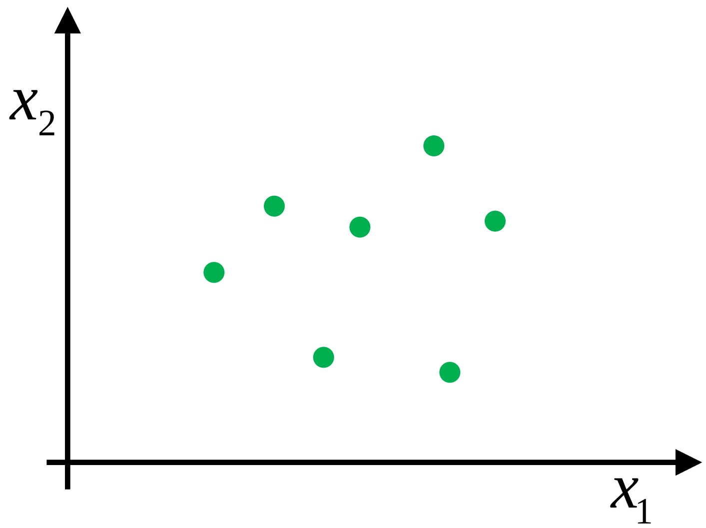
Alternative 2: Embedding



→ Dissimilarity matrix D → X

Training set

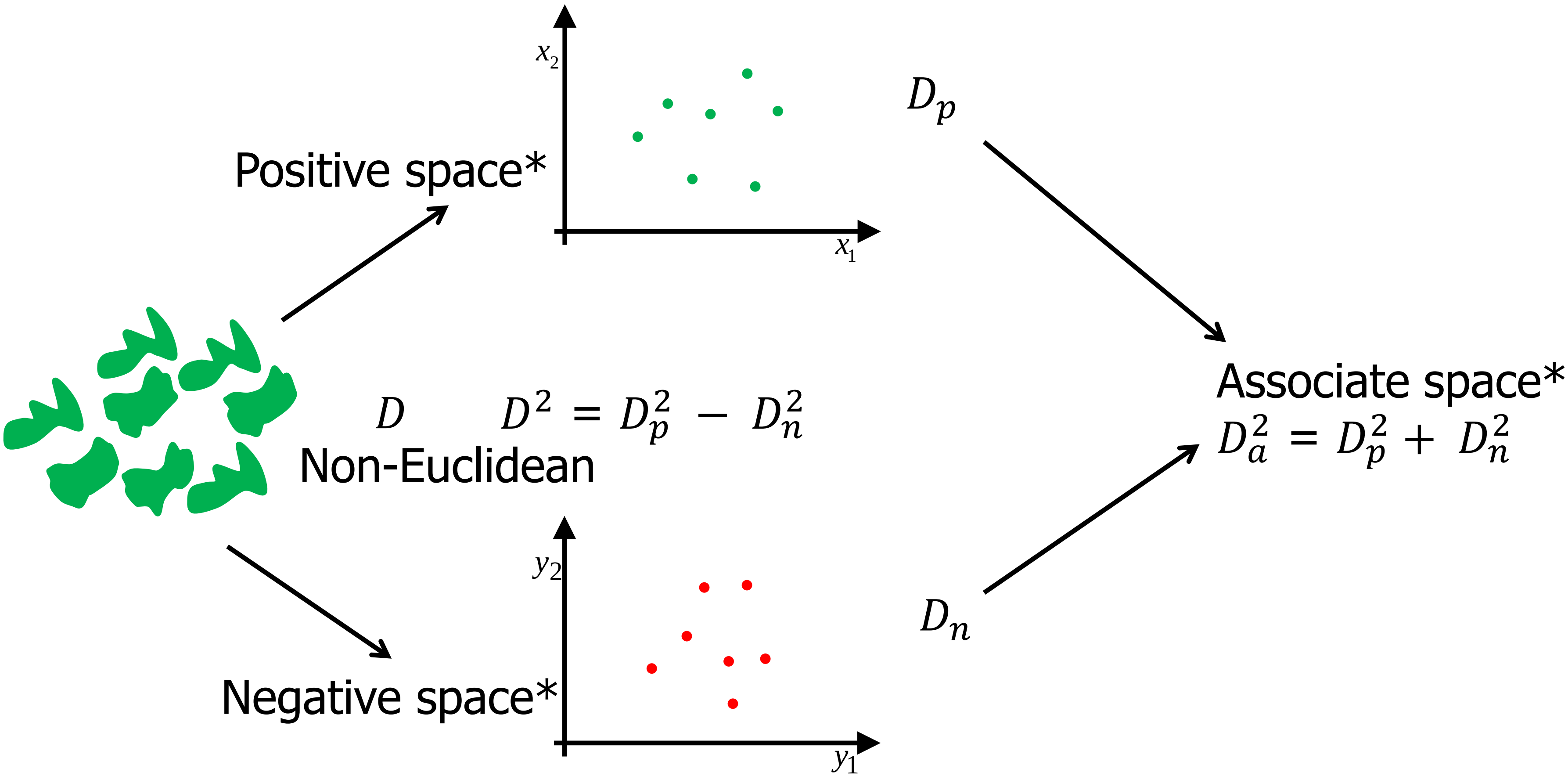
Is there a feature space for which Dist$(X, X) = D$?



Position points in a vector space such that their Euclidean distances → D

Not possible if D is non-Euclidean

Pseudo-Euclidean embedding



* Euclidean

TUDelft

Blob Recognition

446 binary images, varying size, e.g.: 100 x 130

Andreu, G., Crespo, A., Valiente, J.M.: Selecting the toroidal self-organizing feature maps (TSOFM) best organized to object recogn. In: ICNN. (1997) 1341–1346.

Shape classification by weighted-edit distances (Bunke)

*Bunke, H., Buhler, U.: Applications of approximate string matching to 2D shape recognition. Pattern recognition **26** (1993) 1797–1812*

TU Delft

The Chickenpieces dissimilarity matrices

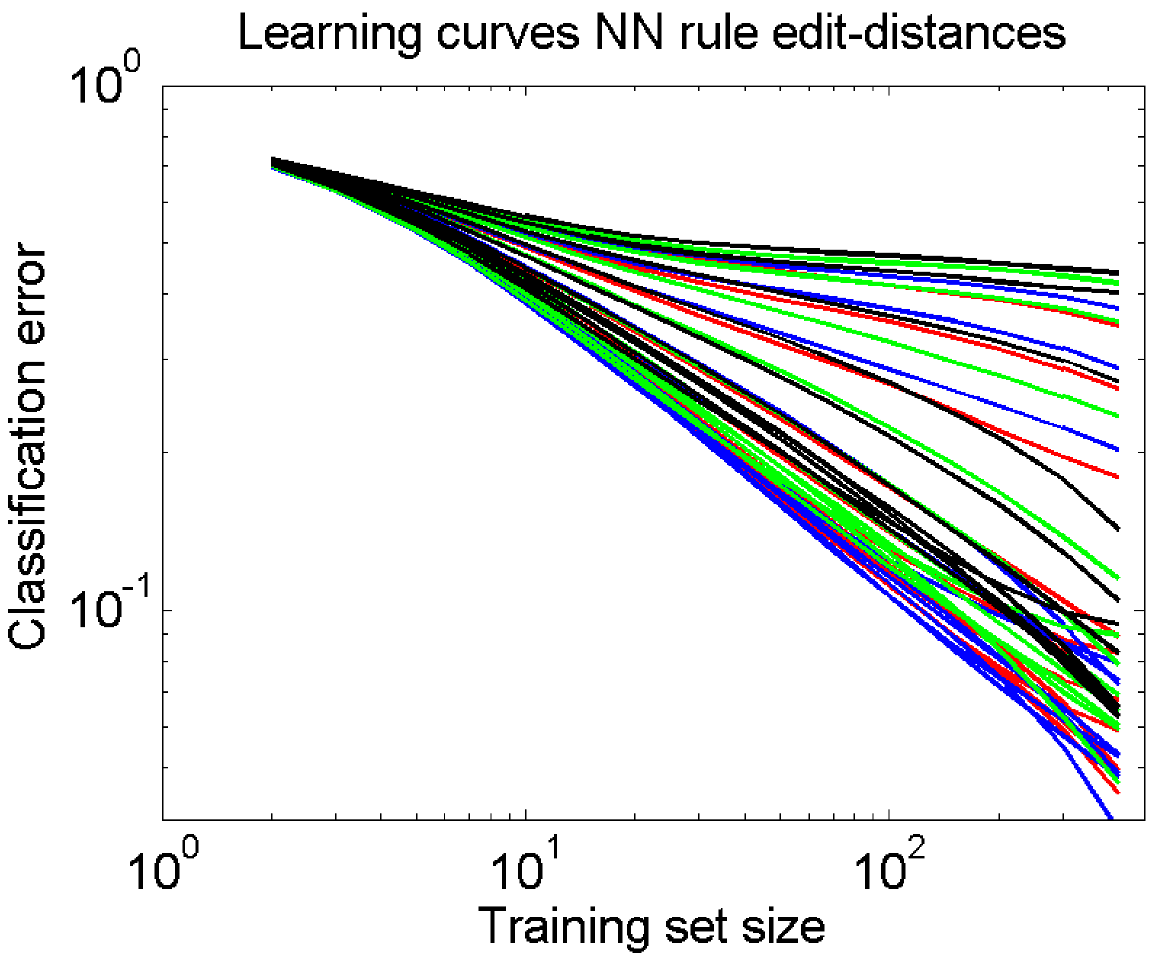
44 Weighted-edit distances measures based on
4 cost functions and
11 string representations.

Shape classification by weighted-edit distances (Bunke)

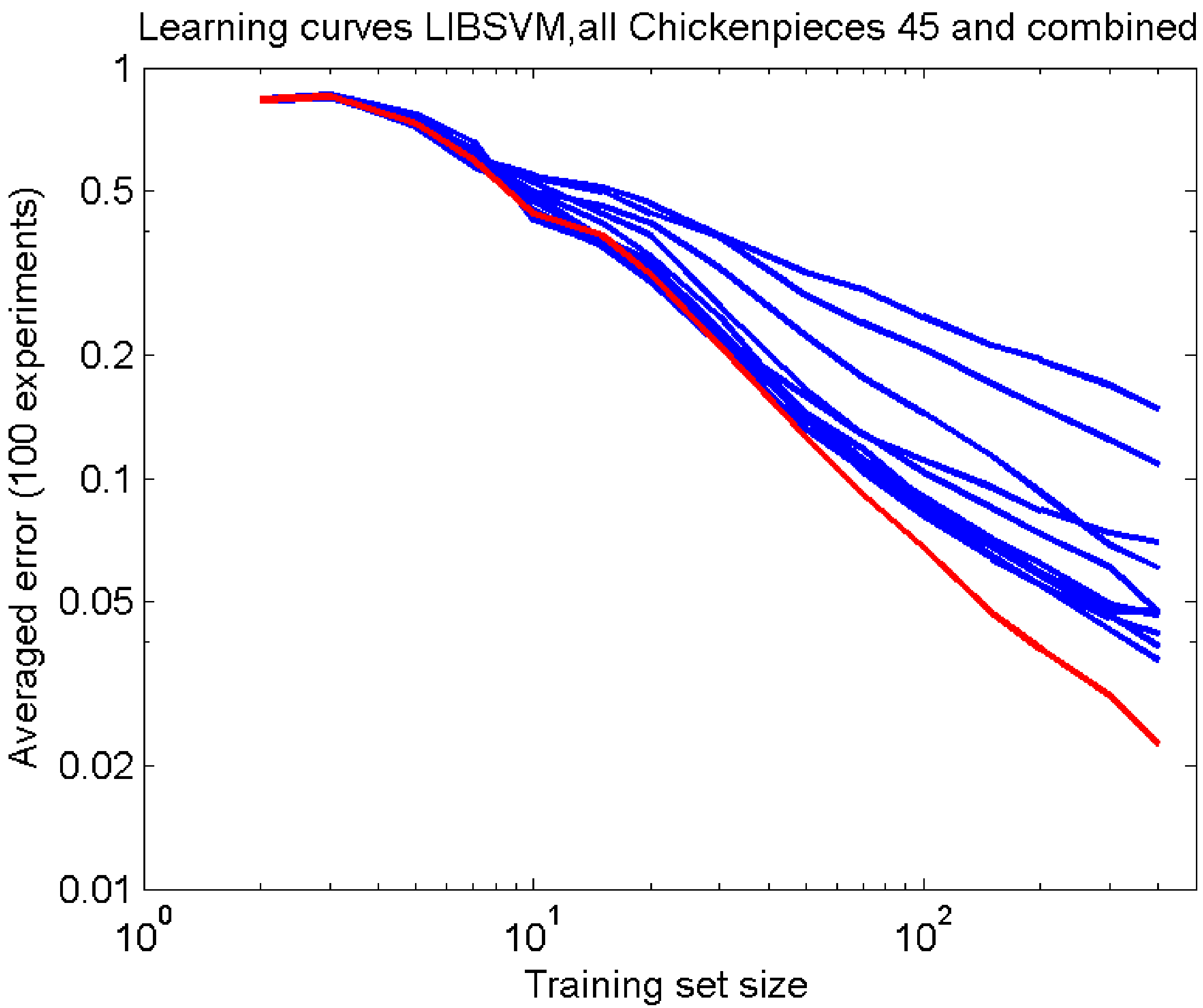
*Bunke, H., Buhler, U.: Applications of approximate string matching to 2D shape recognition. Pattern recognition **26** (1993) 1797–1812*

TU Delft

The Chickenpieces dissimilarity matrices - performances

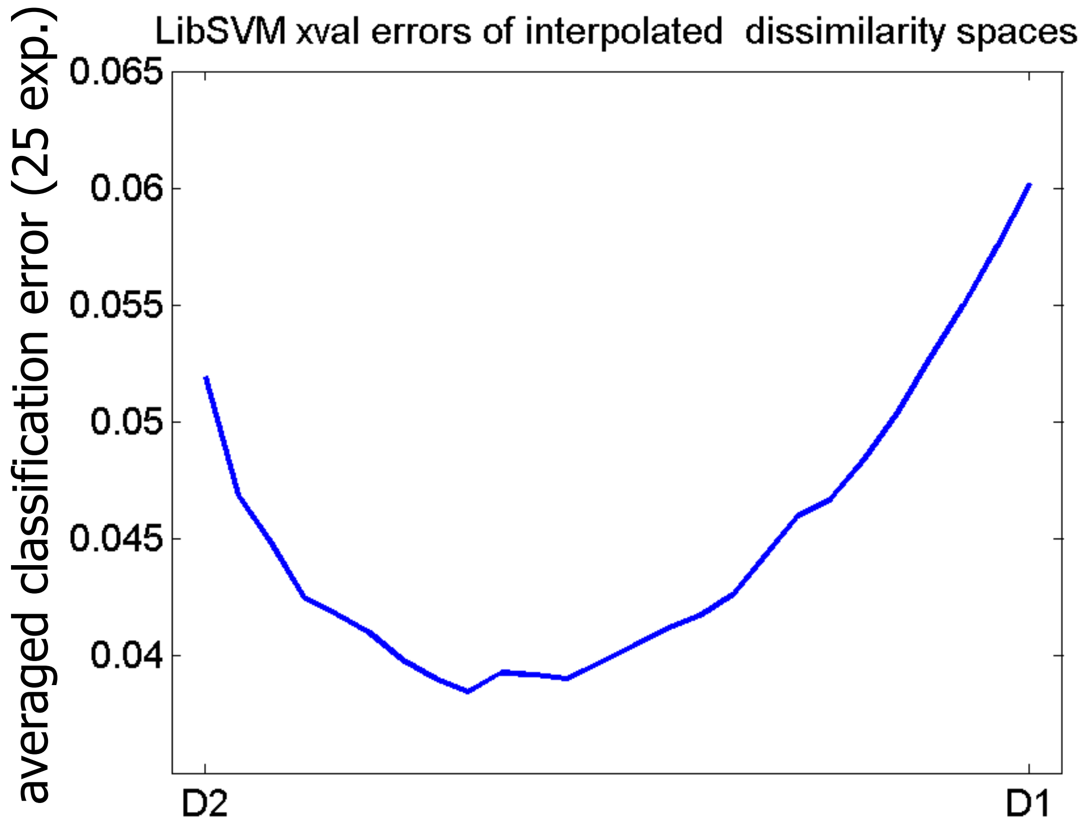


Averaging dissimilarities



Weighted average of two dissimilarity matrices

Chickenpieces-15-45 and Chickenpieces-25-60



Averaging helps!

TU Delft

Chickenpieces: Non-Euclideaness is informative

Average dissimilarity matrices for every cost function: $D_c^2 = \frac{1}{11}\sum_{i=1}^{11} D_{c,i}^2$

Split in Euclidean matrices: $D_c^2 = D_p^2 - D_n^2$

Subtracting helps!

Non-Euclideaness is informative

Cross validation errors

	D_c	D_p	D_n
C=1	0.022	0.137	0.175
C=2	0.020	0.067	0.173
C=3	0.022	0.052	0.148
C=4	0.034	0.108	0.148

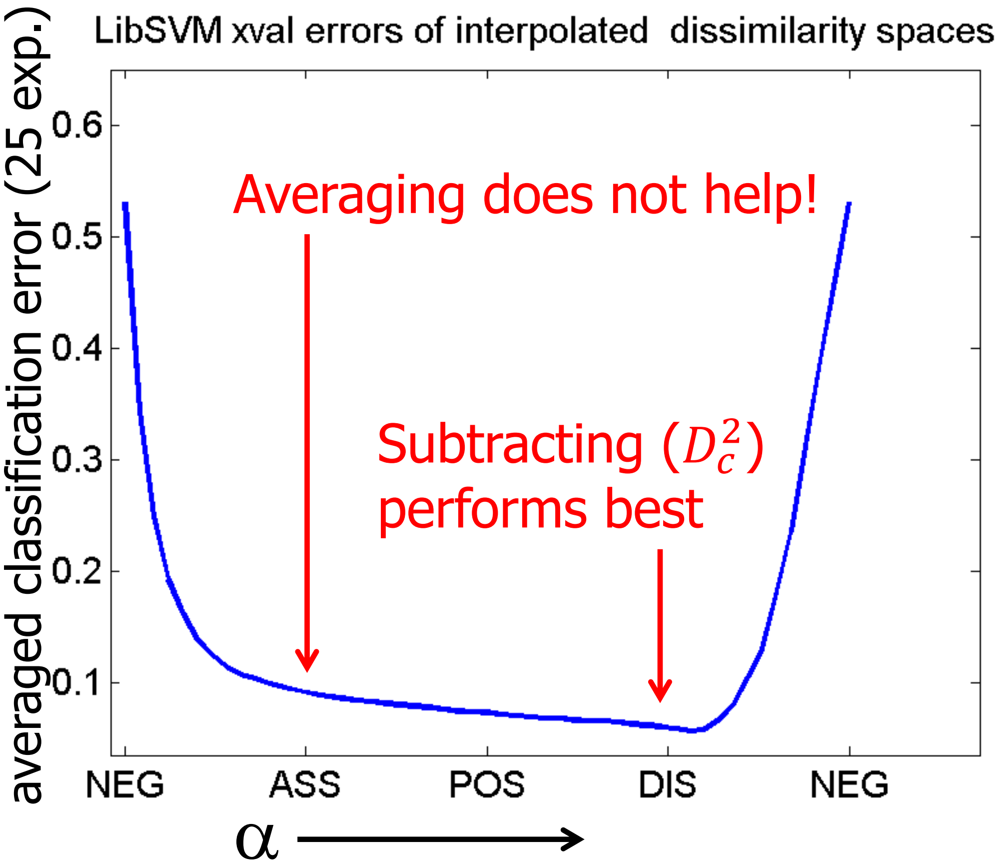
Random assignment error: 0.791

Subtracting dissimilarities

D_c: Chickenpieces-25-60

$D_c^2 = D_p^2 \; - D_n^2$

$D^2 = \sin(\alpha)\, D_p^2 + \cos(\alpha) D_n^2$



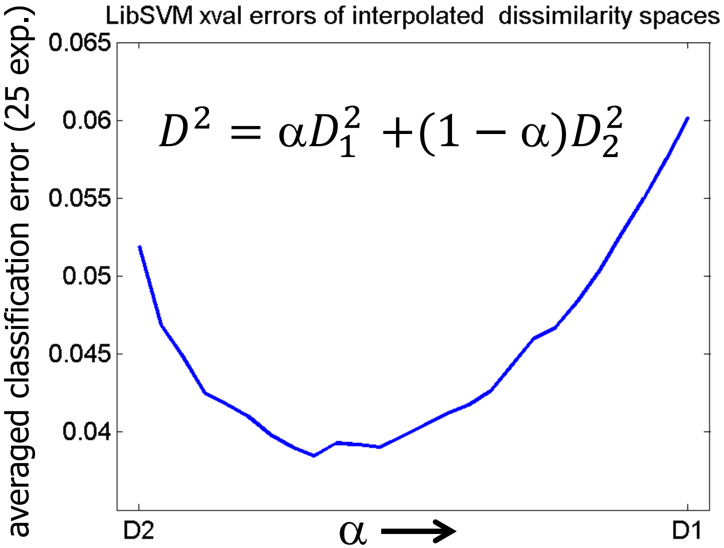
Question

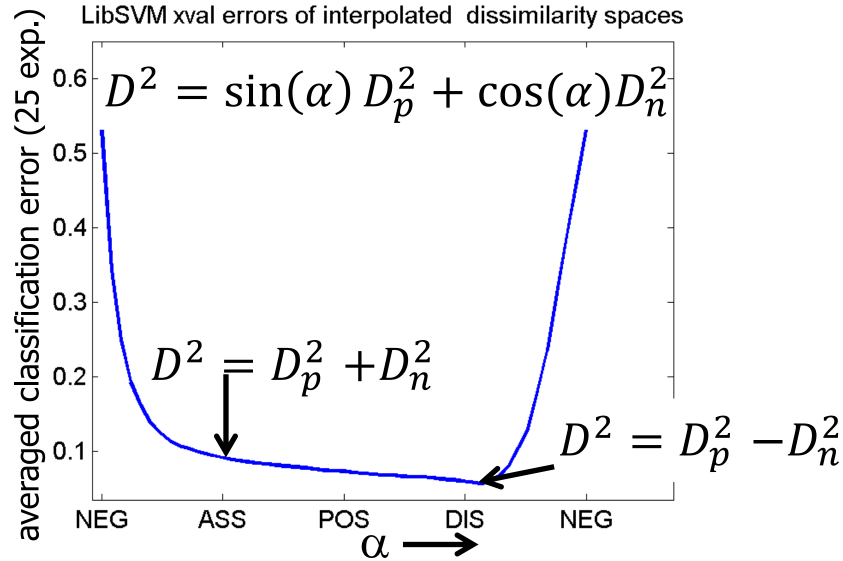
Averaging different (non-Euclidean) dissimilarity measures **may help**.

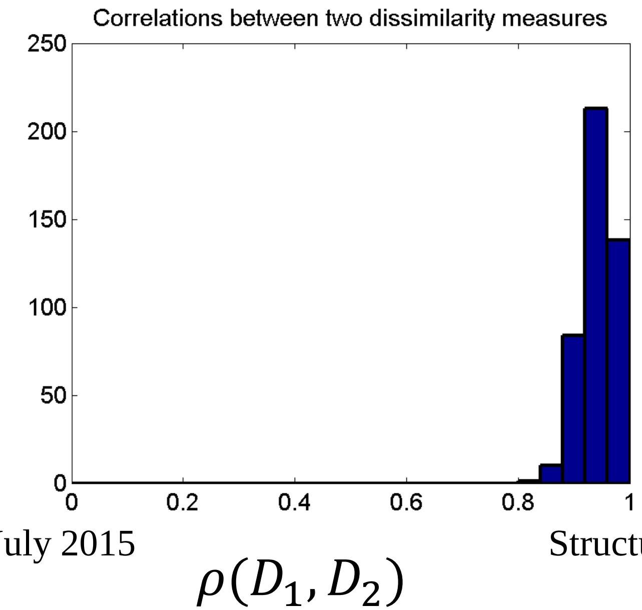
A single non-Euclidean dissimilarity, however, may perform better than the difference of its two constituting Euclidean parts.
So **subtracting may help as well.**

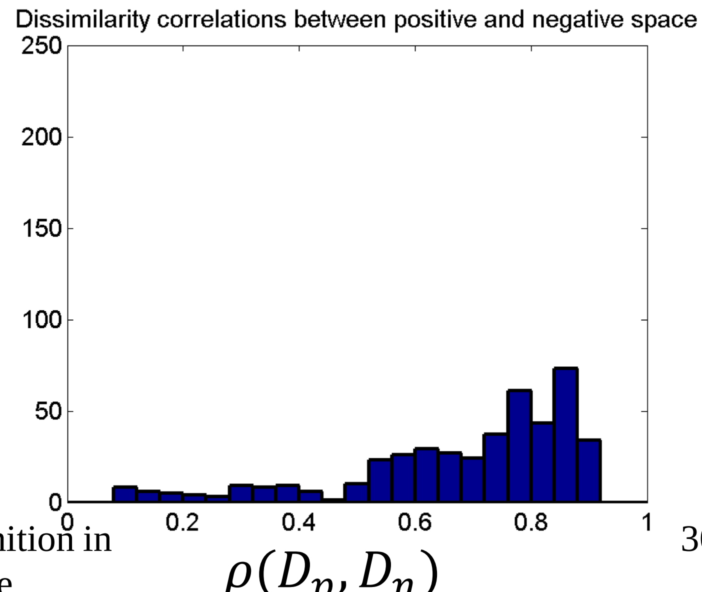
How can we understand this??

Correlations between dissimilarity vectors

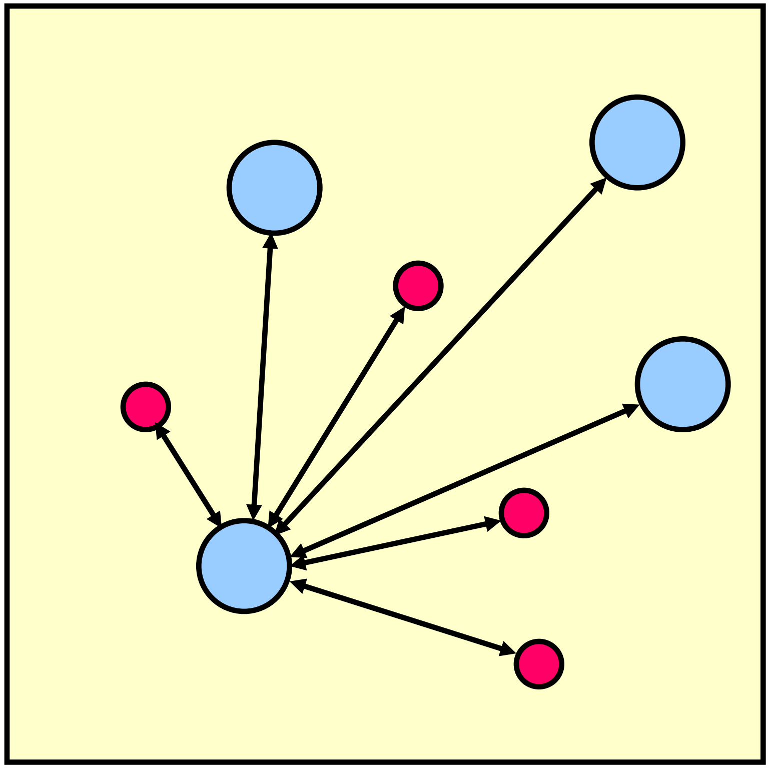








Ball Distances



- Generate sets of balls (classes) uniformly, in a (hyper)cube; not intersecting.
- Balls of the same class have the same size.
- Compute all distances between the ball surfaces.

-> Dissimilarity matrix D

Duin et al., Non-Euclidean dissimilarities: Causes and informativeness, SSSPR 2010, 324-333.

Ball distances: Non-Euclideaness is very informative

$2\ x\ 100$ balls with two sizes.
Given are all Euclidean surface distances.

Split in Euclidean matrices: $D^2 = D_p^2 - D_n^2$

Cross validation errors

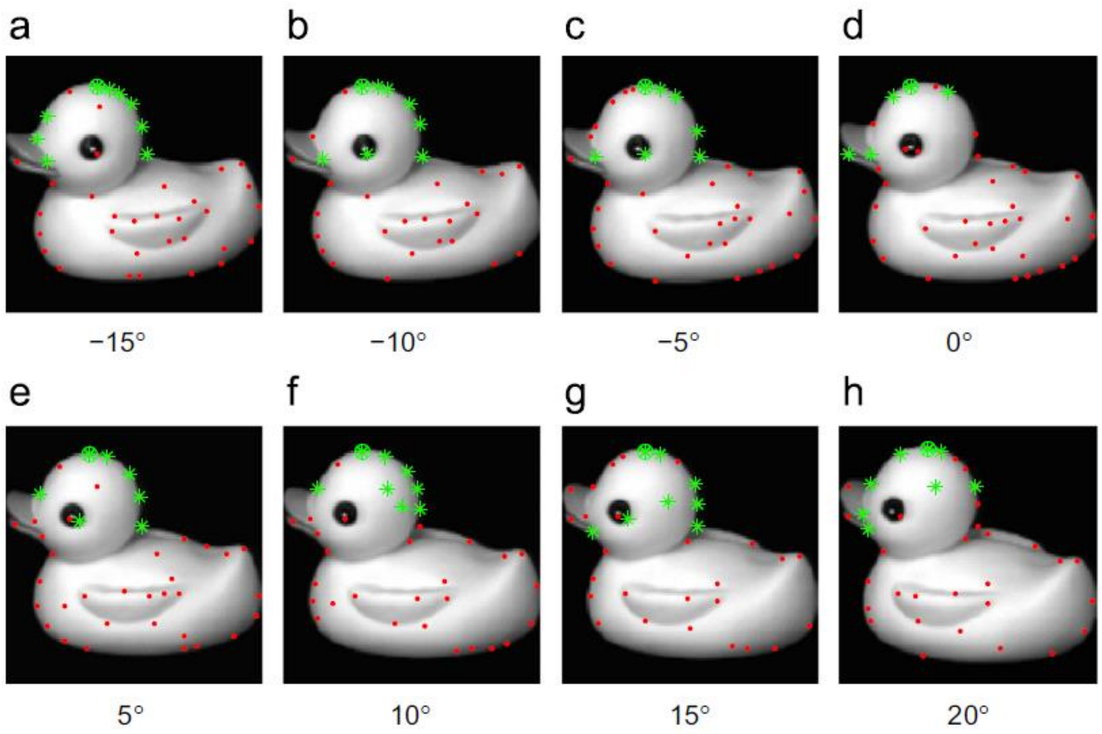
D_c	D_p	D_n
0.470	0.405	0.000

Non-Euclideaness is extremly informative

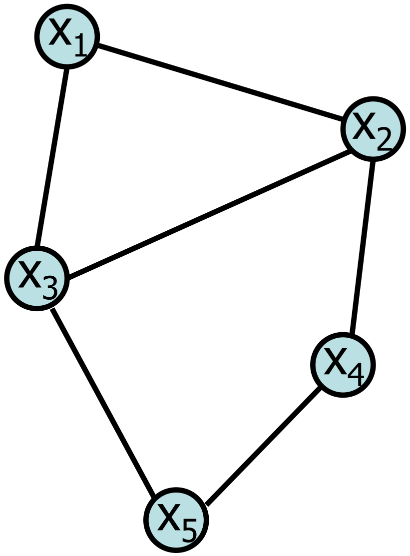
Random assignment error: 0.50

TU Delft

Application: Graphs



Coil dataset



Graph with feature nodes

Taken from: Ren, Aleksic, Wilson, Hancock,
A polynomial characterization of hypergraphs using the Ihara zeta function,
Pattern Recognition, 2011, 1941-1957

Coil dataset (100 classes)

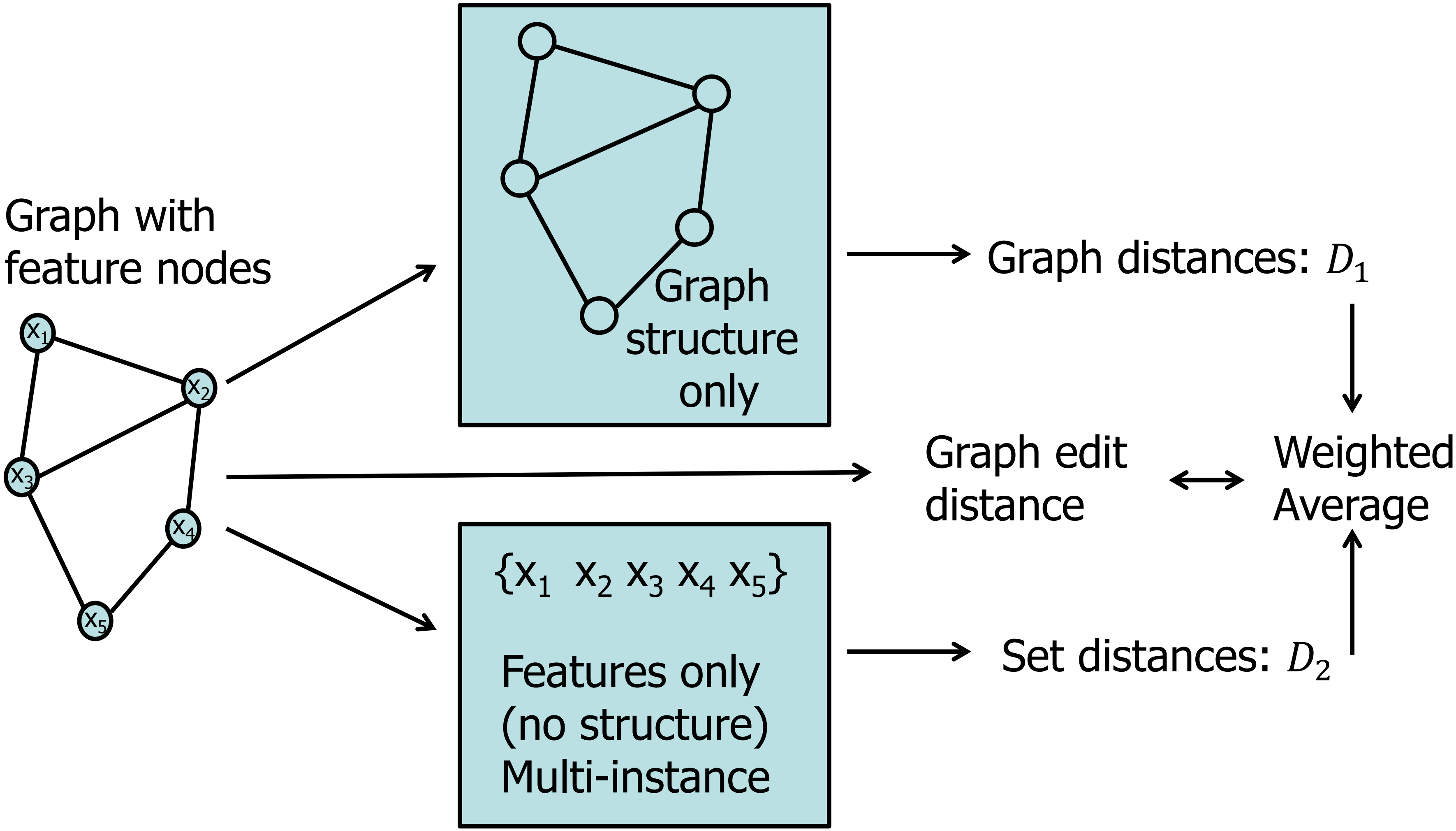
Selection of 10 most difficult classes
72 objects per class

- Segments
- Sift points
- Harris points

3 sets of attributed graphs

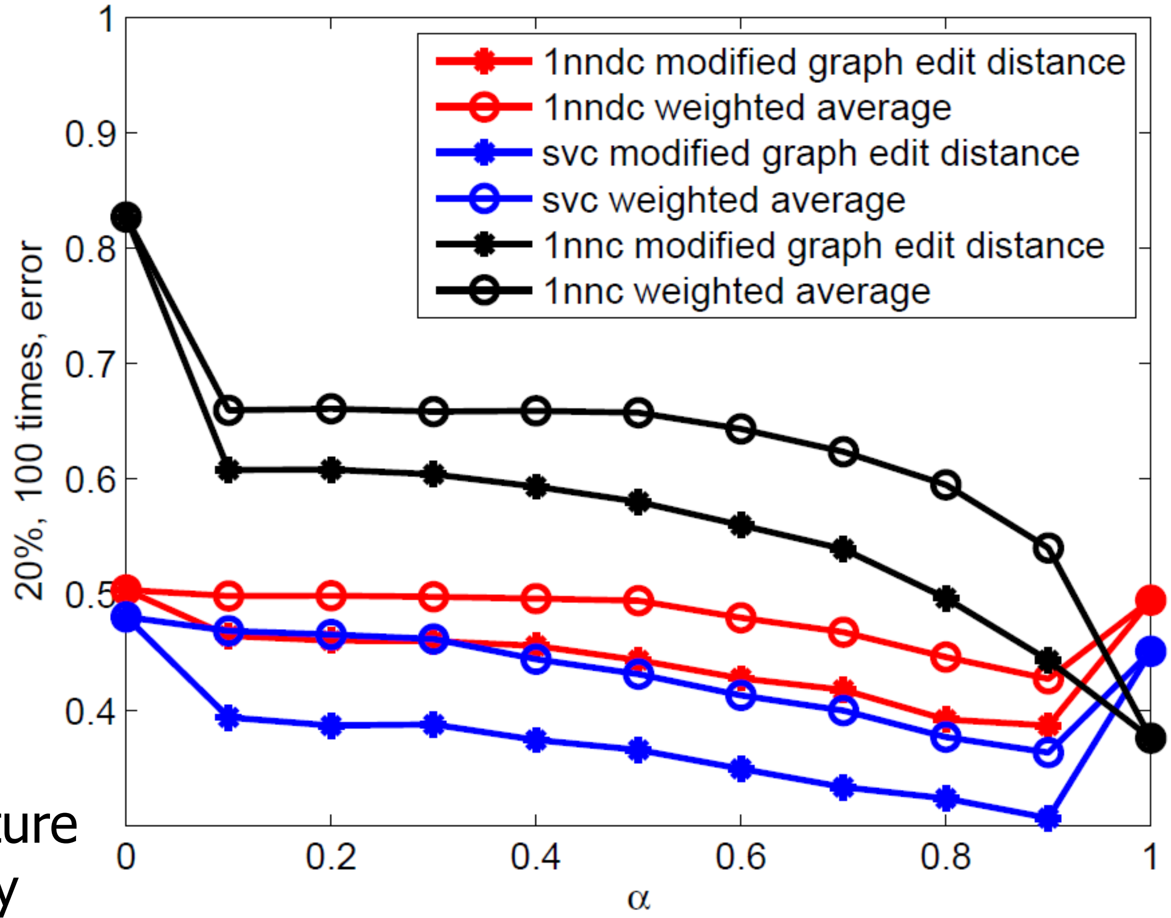
W.R.Lee, V. Cheplygina, D.M.J. Tax, M. Loog, R.P.W. Duin
Bridging Structure and Feature Representations in Graph Matching, IJPRAI,2012

Graphs represented by distance measures



TU Delft

Interpolating structural and feature space dissimilarities

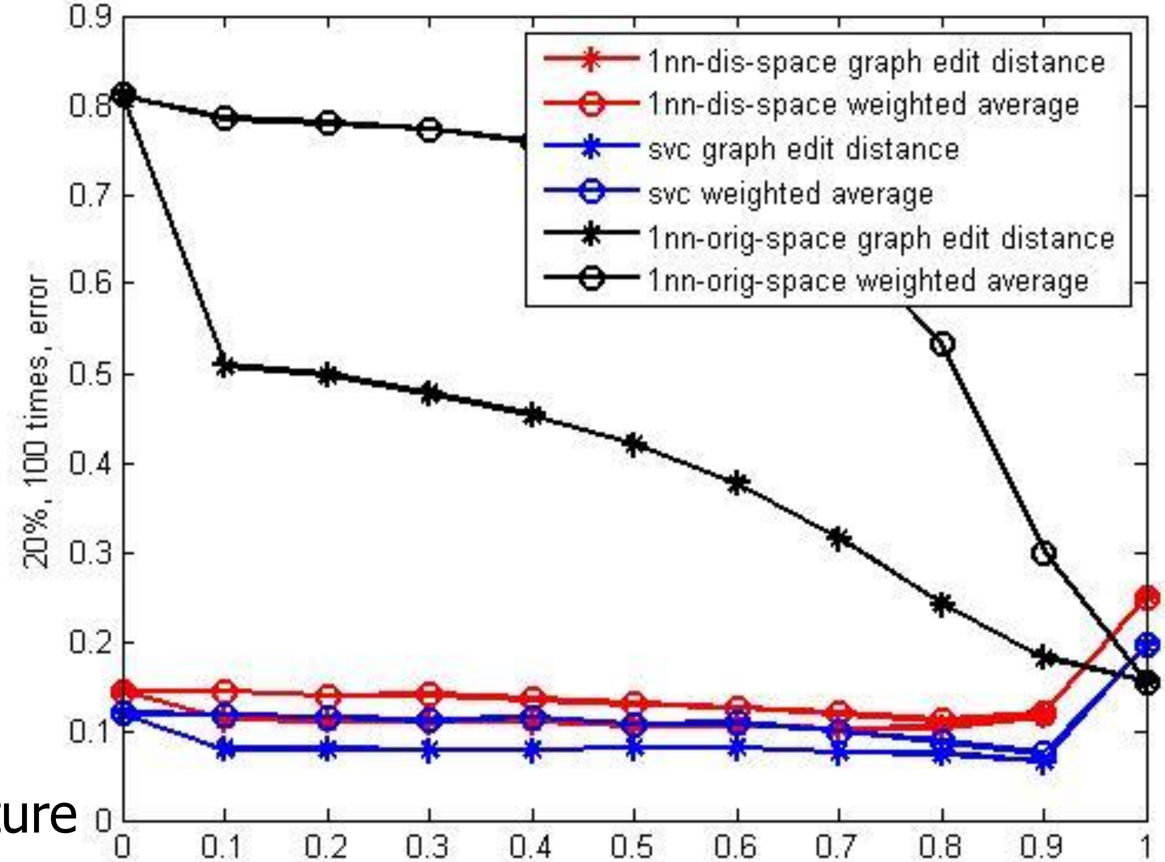


Structure only

Multiple Instance

Segment based graphs

Interpolating structural and feature space dissimilarities



Structure only

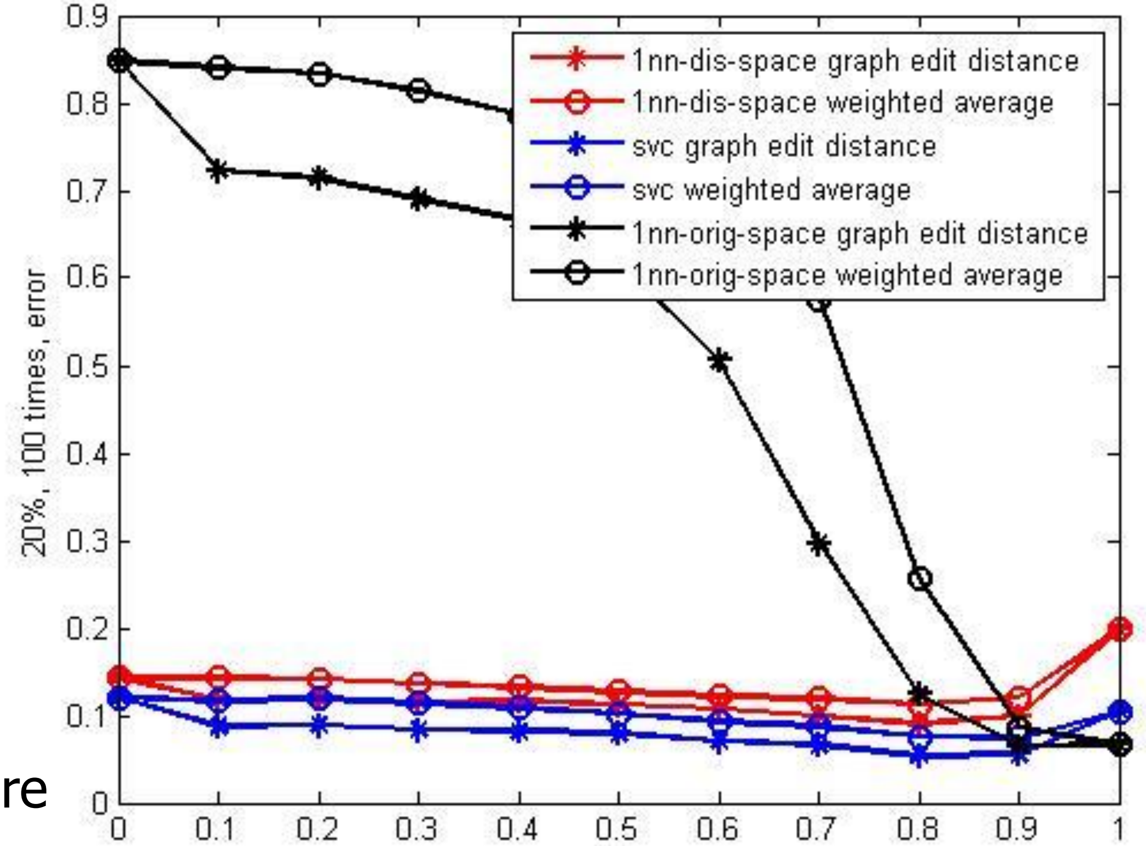
Harris point based graphs

Multiple Instance

TU Delft

Interpolating structural and feature space dissimilarities

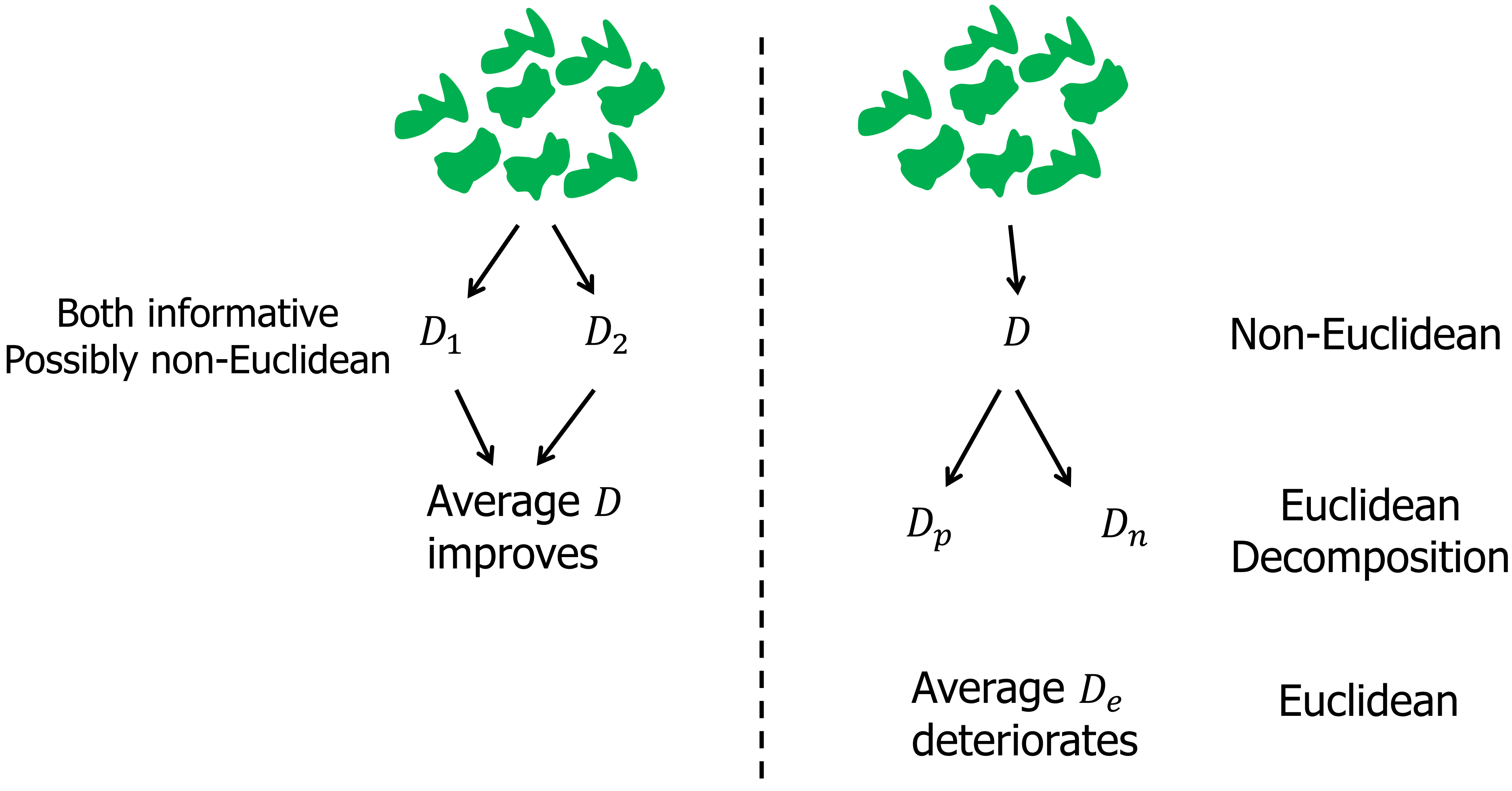
Structure only



Multiple Instance

Sift point based graphs

Observations



TU Delft

Conclusions

- **Combining** dissimilarity representations based on different measures **may improve** the performance
  - by addition as well as by subtraction
  - for Euclidean as well as for non-Euclidean measures

- Combining the positive and negative Euclidean representations obtained by decomposing a non-Euclidean representation improves the performance just by subtraction
  → **Non-Eulideaness is informative**.

- Can we predict the behavior by just studying the representations before combining?

TUDelft