PIF: Anomaly detection via preference embedding

Filippo Leveni*
filippo.leveni@polimi.it

Luca Magri*
luca.magri@polimi.it

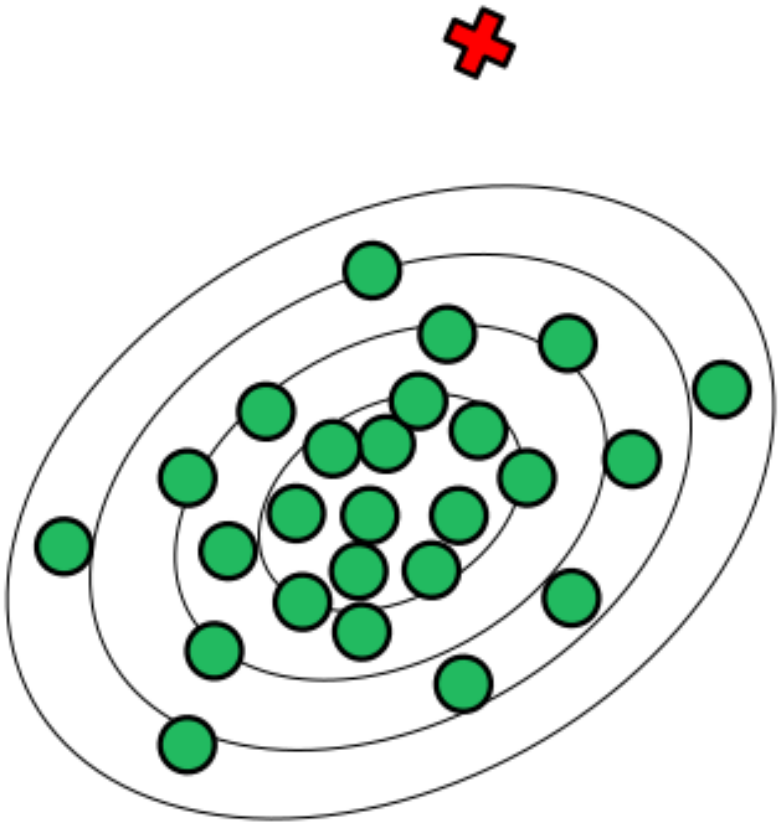
Giacomo Boracchi*
giacomo.boracchi@polimi.it

Cesare Alippi*†
cesare.alippi@polimi.it
cesare.alippi@usi.ch

*Politecnico di Milano, Italy

†Università della Svizzera italiana, Switzerland

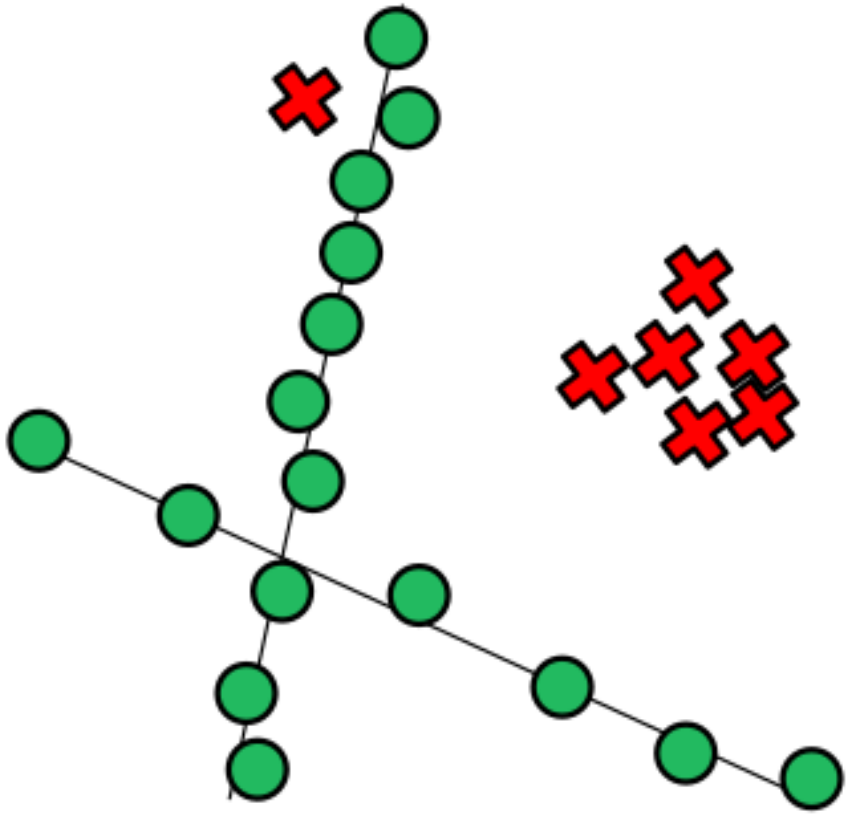
Anomaly detection



Anomaly detection deals with the problem of **identifying data** that **do not conform to an expected behavior**.

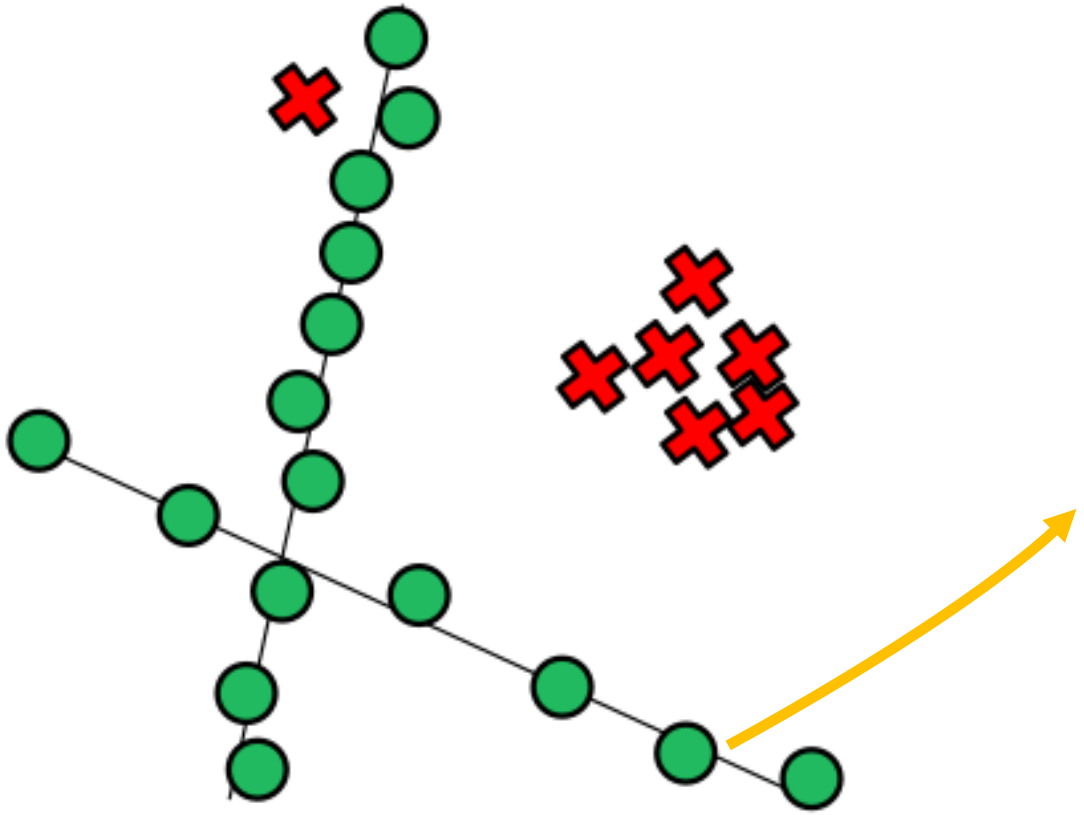
In the statistical and data-mining literature, anomalies are typically detected as **samples falling in low-density regions** of a probability density model describing the data.

Anomaly detection



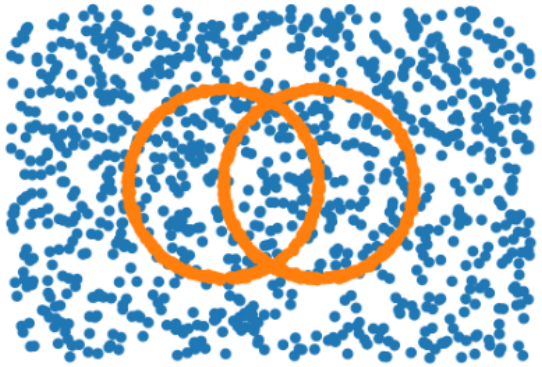
In this work, we consider anomaly detection in a **pattern-recognition** setup, where anomalies are samples that **deviate from unknown structures or patterns**.

Anomaly detection

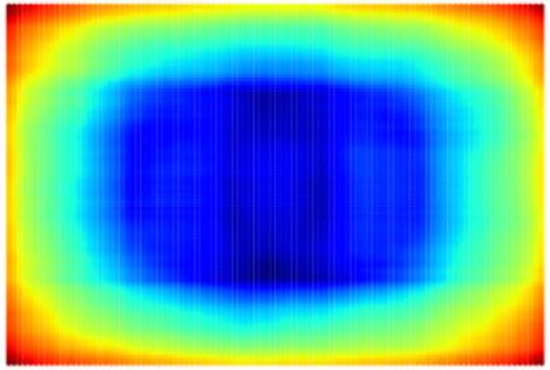


Structures/patterns conform to a family of known geometric primitives.

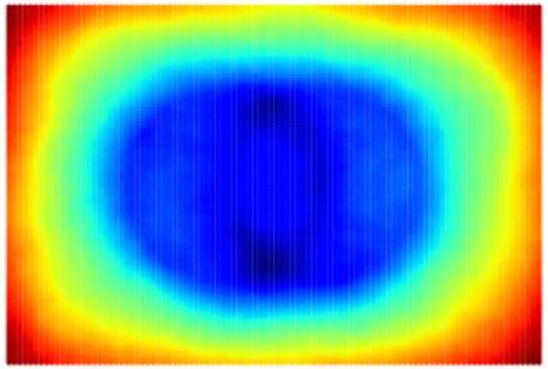
Anomaly detection



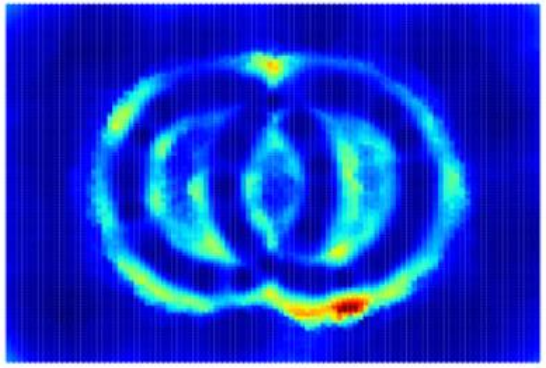
Ground truth



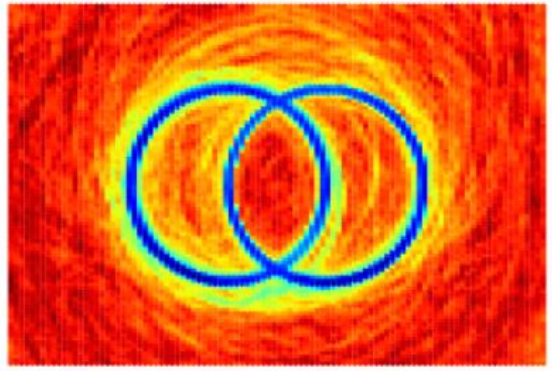
IFOR [1]



EIFOR [2]



LOF [3]



PIF

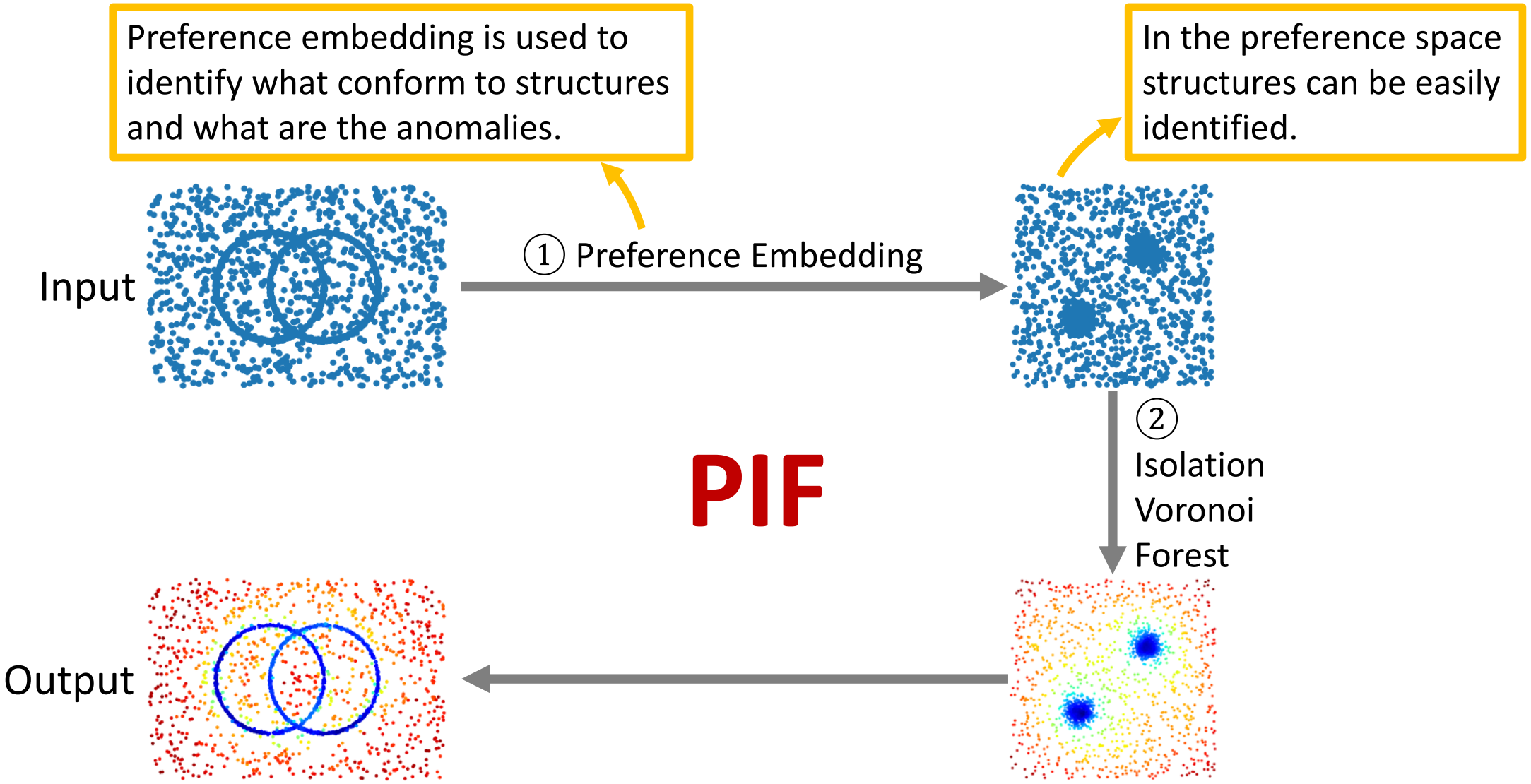
[1] F. T. Liu, K. M. Ting, and Z.-H. Zhou, “Isolation forest”, in International Conference on Data Mining, IEEE, 2008, pp. 413–422.

[2] S. Hariri, M. C. Kind, and R. J. Brunner, “Extended isolation forest”, arXiv preprint arXiv:1811.02141, 2018.

[3] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, “Lof: Identifying density-based local outliers”, in International Conference on Management of data, 2000, pp. 93–104.

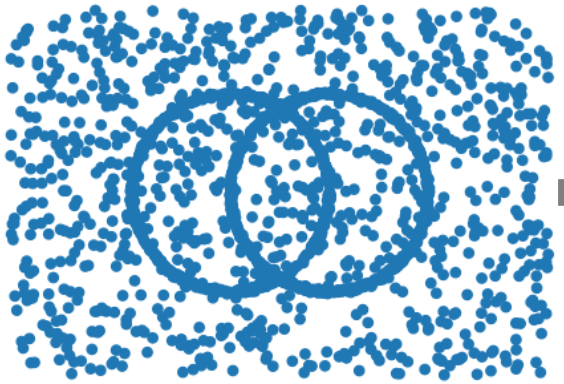
PIF: Our solution

Preference Isolation Forest

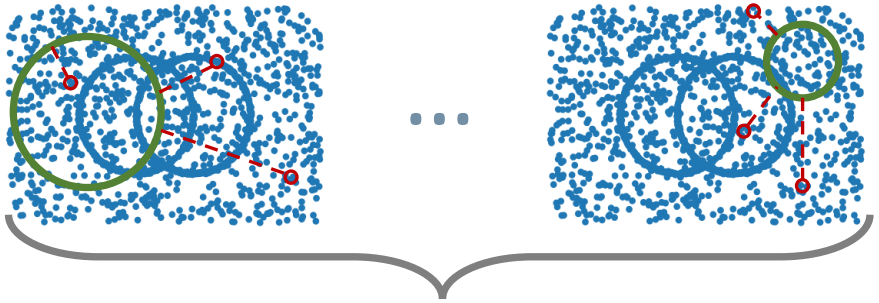

Preference embedding is used to identify what conform to structures and what are the anomalies.
In the preference space structures can be easily identified.
Input
① Preference Embedding
② Isolation Voronoi Forest
PIF
Output

Preference embedding

Initial space

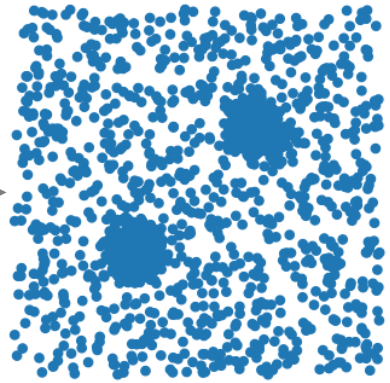


Preference embedding



m model instances

Preference space



$\mathbb{R}^{n\ x\ 2}$

7.2	1.3
5.4	8.0
9.6	0.7
...	...

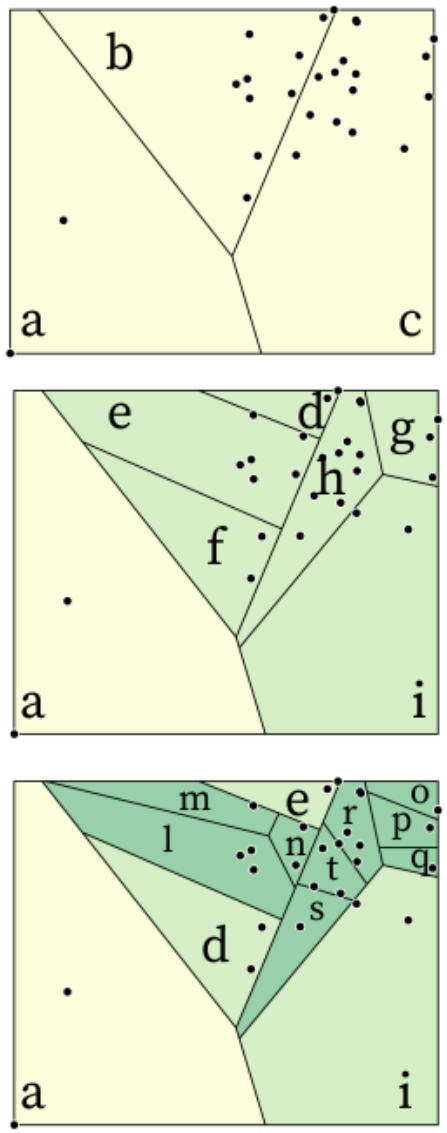
$[0,1]^{n\ x\ m}$

0.6	0.2	...	0.9
0.4	0.8	...	0.6
0.1	0.5	...	0.3
...	...	...	...

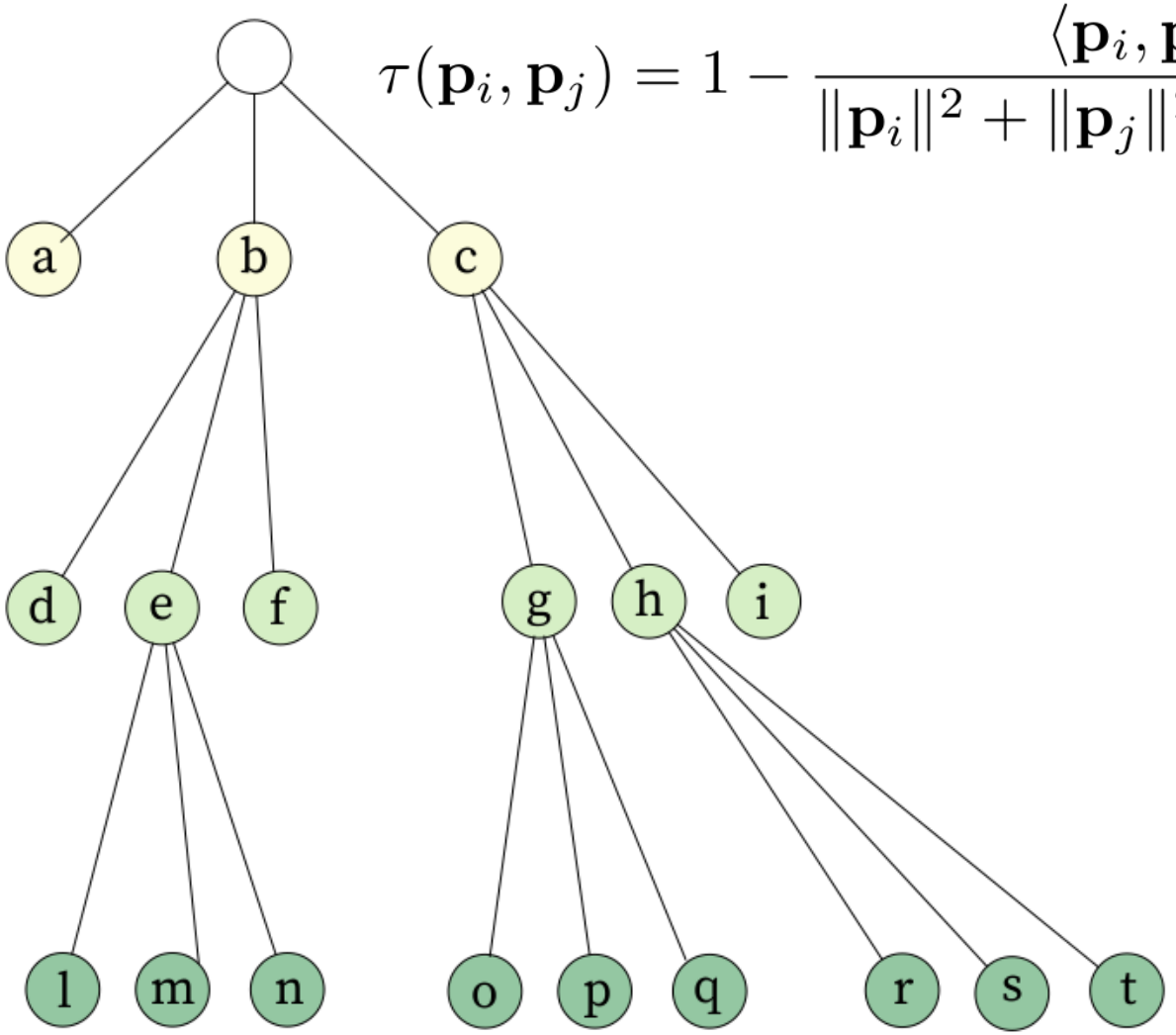
$$[\mathbf{p}_i]_j = \begin{cases} \phi(\delta_{ij}) & \text{if } \delta_{ij} = \mathcal{F}(\mathbf{x}_i, \boldsymbol{\theta}_j) \leq 3\sigma \\ 0 & \text{otherwise} \end{cases}$$

$$\phi(\delta) = \exp(-\delta^2/\sigma)$$

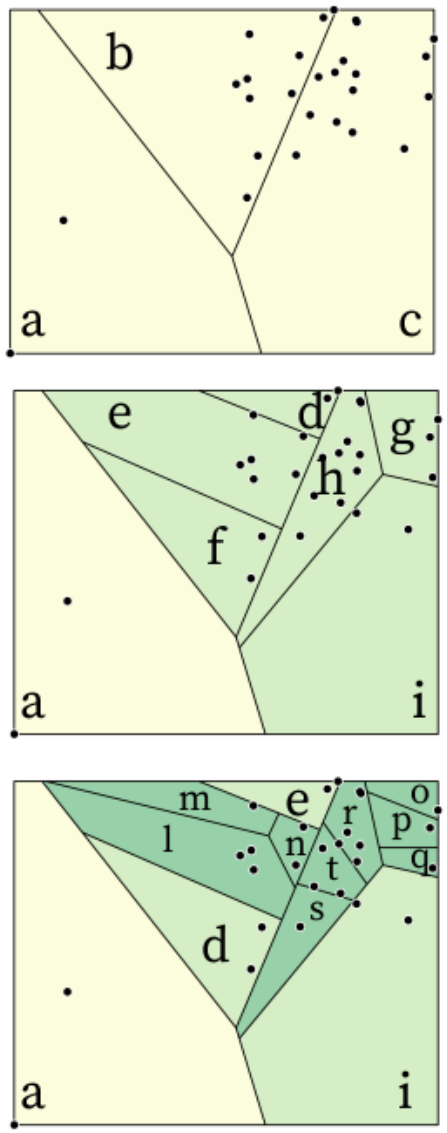
Isolation Voronoi Forest

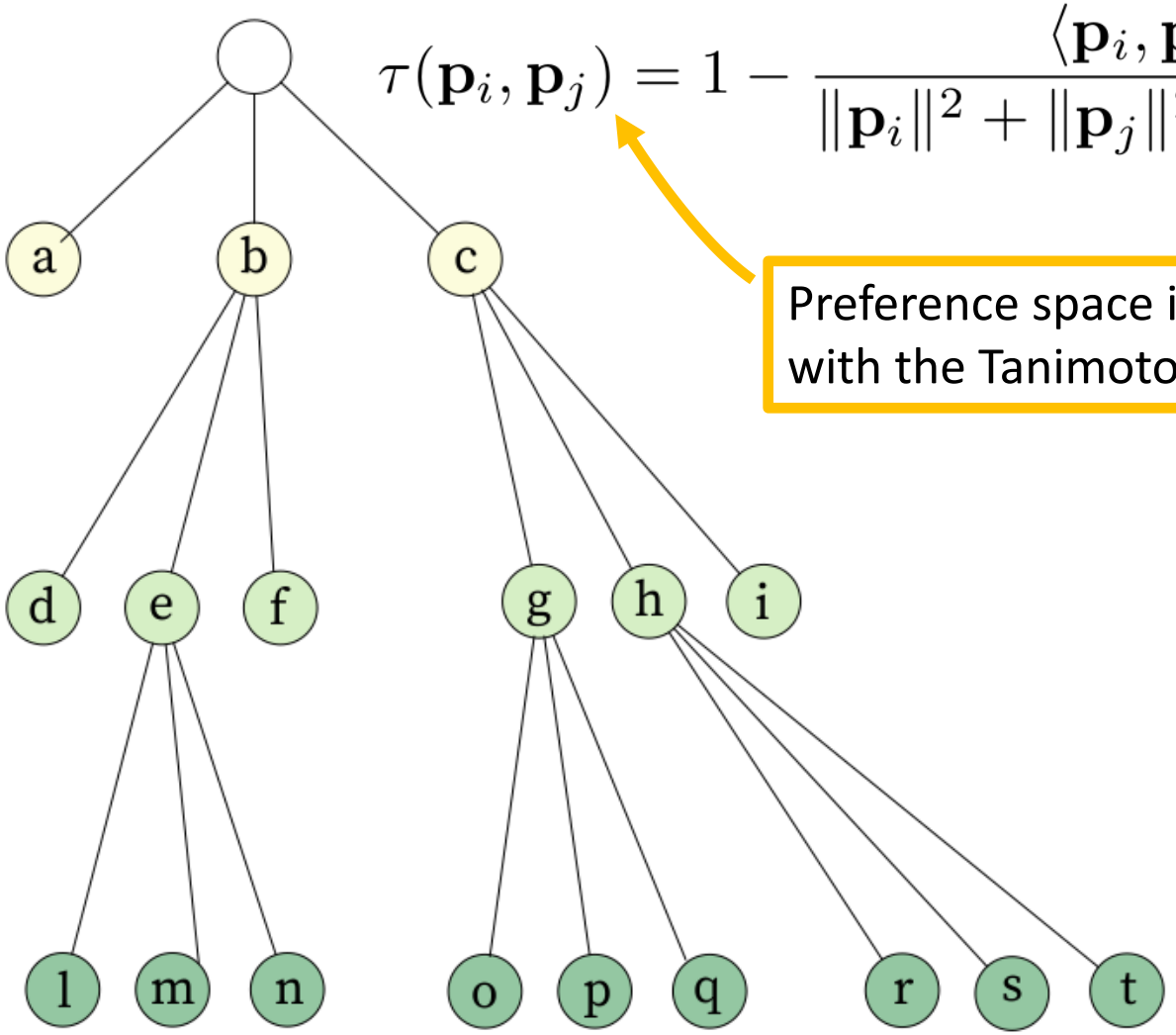


$$\tau(\mathbf{p}_i, \mathbf{p}_j) = 1 - \frac{\langle \mathbf{p}_i, \mathbf{p}_j \rangle}{\|\mathbf{p}_i\|^2 + \|\mathbf{p}_j\|^2 - \langle \mathbf{p}_i, \mathbf{p}_j \rangle}$$



Isolation Voronoi Forest

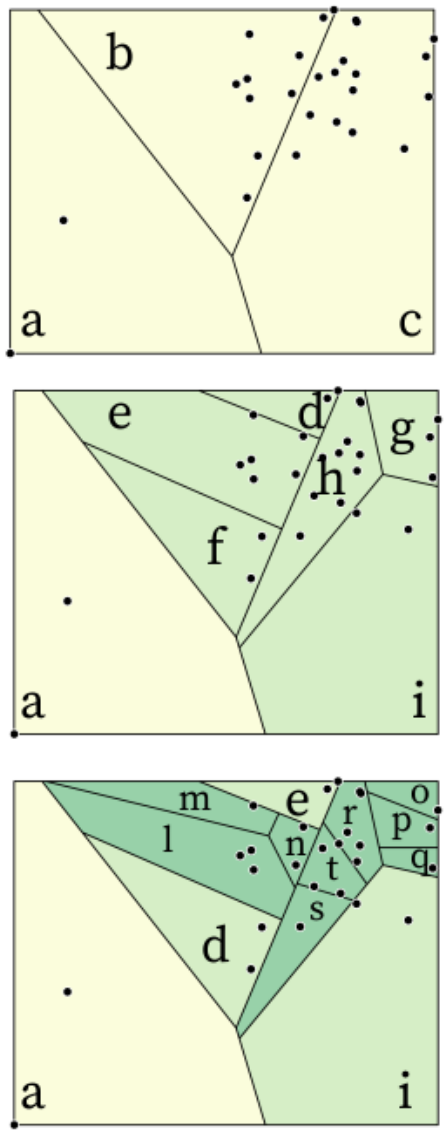




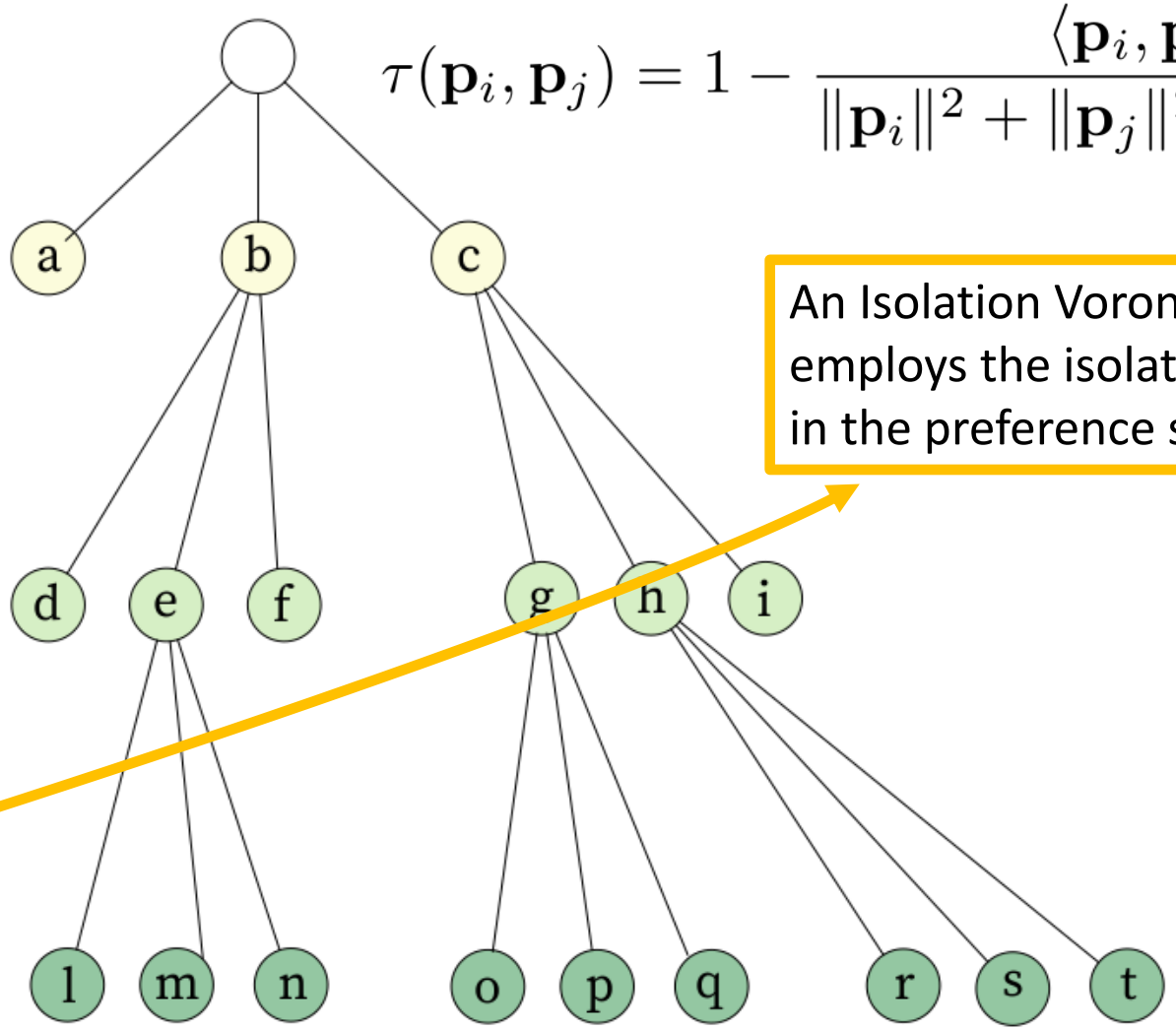
$$\tau(\mathbf{p}_i, \mathbf{p}_j) = 1 - \frac{\langle \mathbf{p}_i, \mathbf{p}_j \rangle}{\|\mathbf{p}_i\|^2 + \|\mathbf{p}_j\|^2 - \langle \mathbf{p}_i, \mathbf{p}_j \rangle}$$

Preference space is equipped with the Tanimoto distance.

Isolation Voronoi Forest



$$\tau(\mathbf{p}_i, \mathbf{p}_j) = 1 - \frac{\langle \mathbf{p}_i, \mathbf{p}_j \rangle}{\|\mathbf{p}_i\|^2 + \|\mathbf{p}_j\|^2 - \langle \mathbf{p}_i, \mathbf{p}_j \rangle}$$

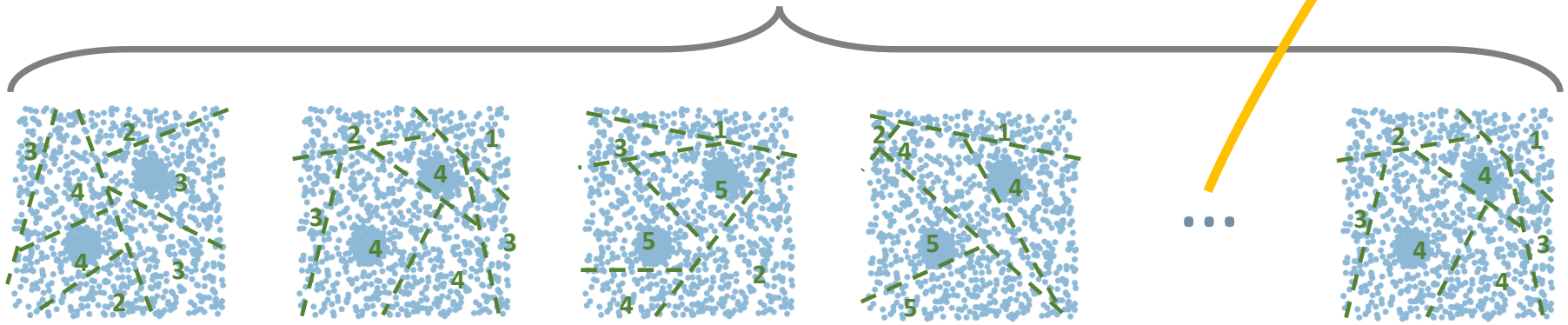


An Isolation Voronoi Tree employs the isolation principle in the preference space.

Isolation Voronoi Forest

Isolation Voronoi Forest is very efficient.

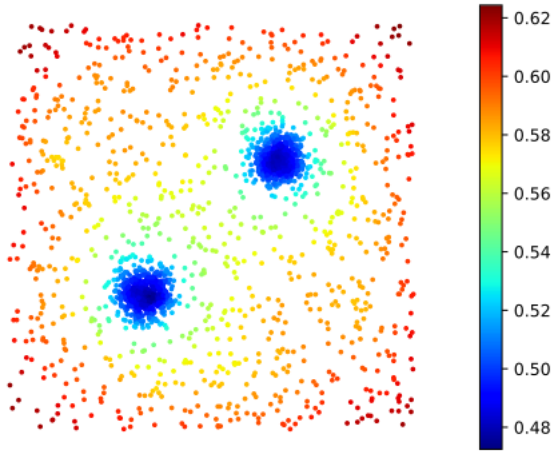
k nested Voronoi tessellations

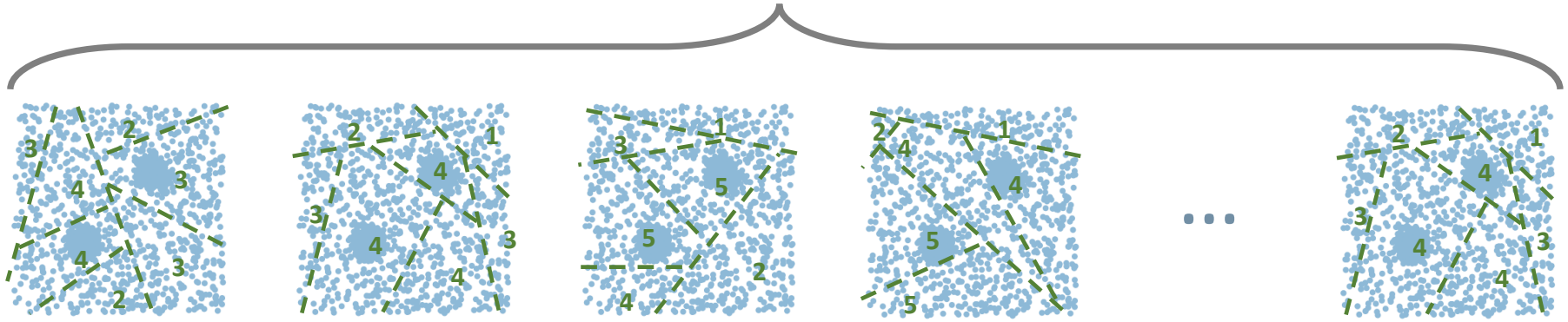


Isolation Voronoi Forest

$$\alpha_\psi(\mathbf{p}) = 2^{-\frac{E(\mathbf{h}(\mathbf{p}))}{c(\psi)}}$$

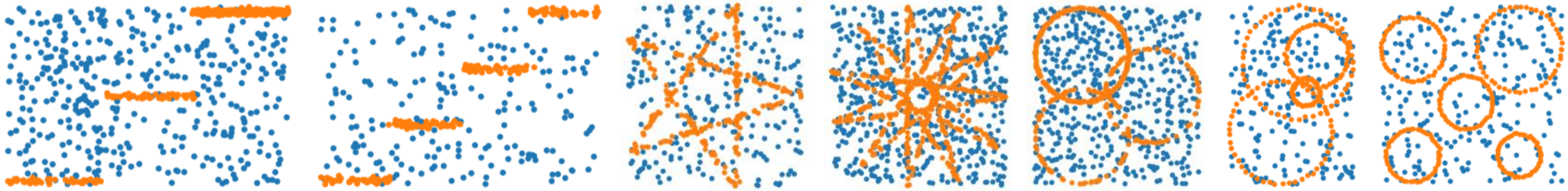
$$c(n) = \begin{cases} 0 & \text{if } n = 1 \\ 1 & \text{if } n = 2 \\ 2H(n-1) - 2(n-1)/n & \text{if } n > 2 \end{cases}$$





Results

Synthetic datasets

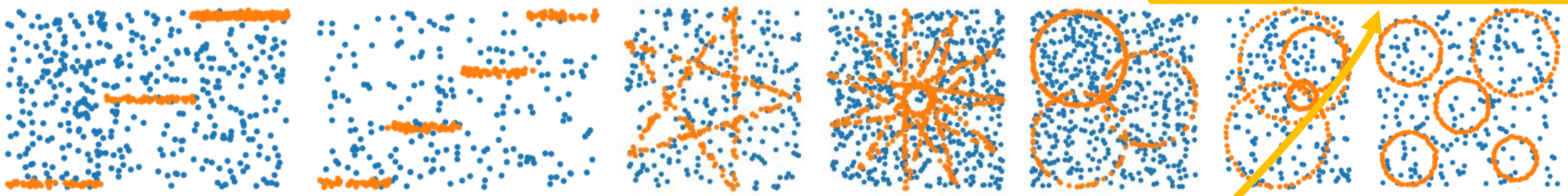


stair3 stair4 star5 star11 circle3 circle4 circle5

	Euclidean				Preference binary				Preference			
	LOF ℓ_2	IFOR	EIFOR	PIF ℓ_2	LOF jac	IFOR	EIFOR	PIF jac	LOF tani	IFOR	EIFOR	PIF
stair3	0.737	0.925	0.920	0.918	0.904	0.885	0.864	0.958	0.815	0.923	0.925	**0.971**
stair4	0.814	0.889	0.874	0.871	0.849	0.855	0.860	0.941	0.881	0.912	0.908	**0.952**
star5	0.771	0.722	0.738	0.788	0.875	0.745	0.769	0.872	**0.929**	0.761	0.822	0.910
star11	0.671	0.728	0.727	0.738	0.830	0.739	0.741	0.771	**0.900**	0.738	0.774	0.796
circle3	0.761	0.698	0.732	0.779	0.719	0.842	0.854	0.900	0.731	0.854	0.891	**0.930**
circle4	0.640	0.641	0.665	0.679	0.827	0.686	0.699	0.860	0.906	0.667	0.720	0.897
circle5	0.543	0.569	0.570	0.633	0.699	0.597	0.617	0.672	**0.823**	0.573	0.593	0.780
Mean	0.705	0.739	0.747	0.772	0.815	0.764	0.772	0.853	0.855	0.775	0.805	**0.891**

Synthetic datasets

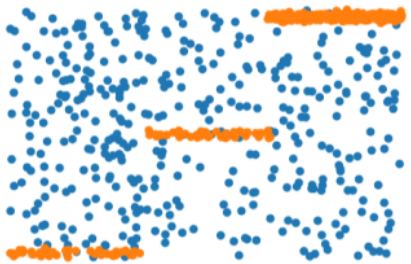
Preference embedding increases the separability between structured and unstructured data.



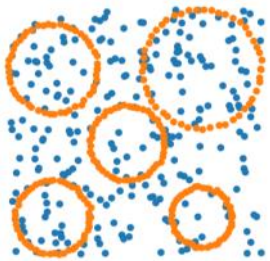
stair3 stair4 star5 star11 circle3 circle4 circle5

	Euclidean				Preference binary				Preference			
	LOF ℓ_2	IFOR	EIFOR	PIF ℓ_2	LOF jac	IFOR	EIFOR	PIF jac	LOF tani	IFOR	EIFOR	PIF
stair3	0.737	0.925	0.920	0.918	0.904	0.885	0.864	0.958	0.815	0.923	0.925	**0.971**
stair4	0.814	0.889	0.874	0.871	0.849	0.855	0.860	0.941	0.881	0.912	0.908	**0.952**
star5	0.771	0.722	0.738	0.788	0.875	0.745	0.769	0.872	**0.929**	0.761	0.822	0.910
star11	0.671	0.728	0.727	0.738	0.830	0.739	0.741	0.771	**0.900**	0.738	0.774	0.796
circle3	0.761	0.698	0.732	0.779	0.719	0.842	0.854	0.900	0.731	0.854	0.891	**0.930**
circle4	0.640	0.641	0.665	0.679	0.827	0.686	0.699	0.860	0.906	0.667	0.720	0.897
circle5	0.543	0.569	0.570	0.633	0.699	0.597	0.617	0.672	**0.823**	0.573	0.593	0.780
Mean	0.705	0.739	0.747	0.772	0.815	0.764	0.772	0.853	0.855	0.775	0.805	**0.891**

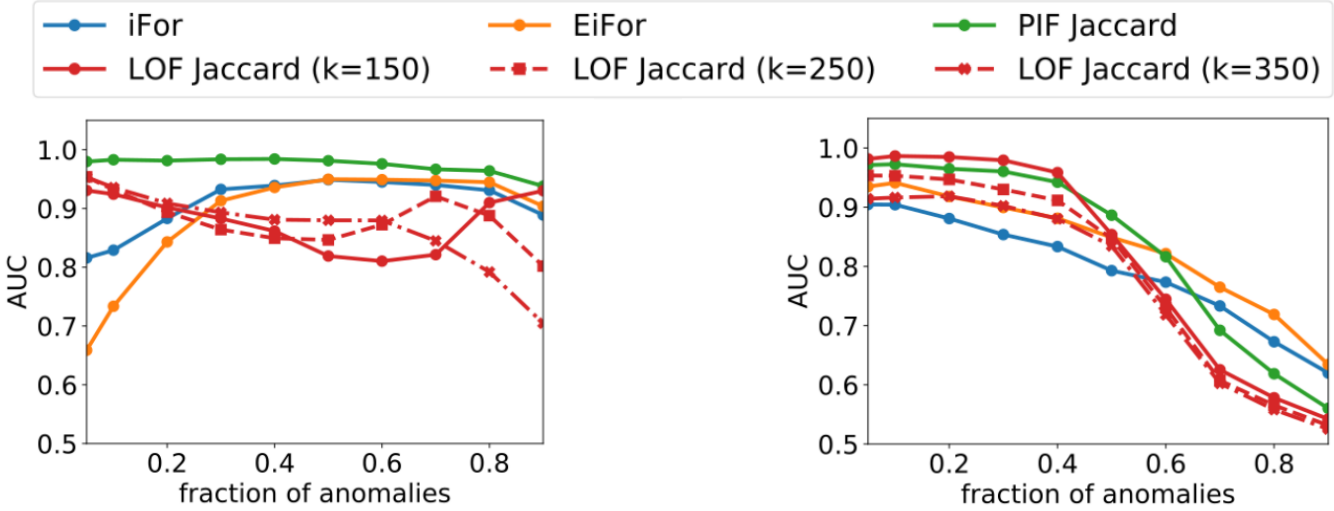
Synthetic datasets



stair3



circle5

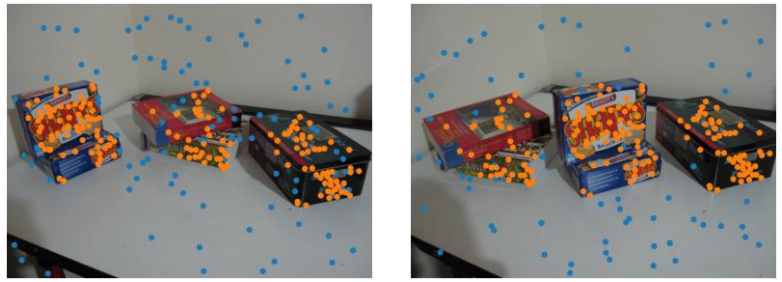


stair3
unbalanced structures

circle5
balanced structures

Real datasets

johnsona



biscuitbookbox

	LOF tani	iFor	EiFor	PIF
barrsmith	**0.969**	0.708	0.715	0.944
bonhall	0.918	0.969	0.967	0.949
bonython	**0.978**	0.679	0.691	0.954
elderhalla	0.999	0.925	0.909	**0.999**
elderhallb	0.986	0.924	0.943	**0.999**
hartley	0.963	0.749	0.793	**0.989**
johnsona	0.993	0.993	0.993	**0.998**
johnsonb	0.776	0.999	0.998	0.999
ladysymon	0.847	0.944	0.943	**0.997**
library	**1.000**	0.764	0.771	0.998
napiera	0.975	0.869	0.879	**0.983**
napierb	0.888	0.931	0.936	**0.953**
neem	0.985	0.896	0.906	**0.996**
nese	**0.996**	0.888	0.892	0.980
oldclassicswing	0.936	0.923	0.943	**0.987**
physics	0.670	0.858	0.787	**1.000**
sene	**0.997**	0.698	0.731	0.988
unihouse	0.785	0.998	0.998	**0.999**
unionhouse	**0.987**	0.639	0.664	0.968
Mean	0.929	0.861	0.866	**0.983**

Homographies

	LOF tani	iFor	EiFor	PIF
biscuit	0.976	0.994	0.996	**1.000**
biscuitbook	1.000	0.987	0.988	1.000
biscuitbookbox	**1.000**	0.990	0.989	0.996
boardgame	**0.962**	0.400	0.304	0.949
book	0.996	1.000	1.000	1.000
breadcartoychips	**0.989**	0.978	0.971	0.976
breadcube	**1.000**	0.998	0.998	0.999
breadcubechips	**0.999**	0.985	0.985	0.998
breadtoy	0.984	0.999	0.998	0.999
breadtoycar	**0.998**	0.933	0.883	0.991
carchipscube	**0.993**	0.981	0.966	0.987
cube	0.999	0.970	0.982	0.999
cubebreadtoychips	0.990	0.962	0.958	0.989
cubechips	**1.000**	0.995	0.994	1.000
cubetoy	**1.000**	0.997	0.995	1.000
dinobooks	0.887	0.873	0.857	**0.899**
game	**1.000**	0.901	0.895	0.999
gamebiscuit	1.000	0.985	0.988	**1.000**
toycubecar	**0.973**	0.290	0.192	0.964
Mean	0.987	0.906	0.891	0.987

Fundamental matrices

H. S. Wong, T.-J. Chin, J. Yu, and D. Suter, “Dynamic and hierarchical multi-structure geometric model fitting”, in International Conference on Computer Vision, 2011

Conclusions

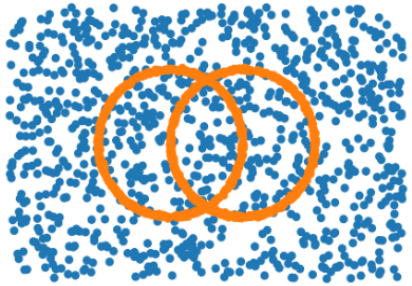
- Empirical evaluation demonstrated that preference embedding **increases the separability** between normal (**structured**) and anomalous (**unstructured**) data.

- PIF **outperforms all the alternatives** where anomaly-detection methods are straightforwardly plugged in the preference space.

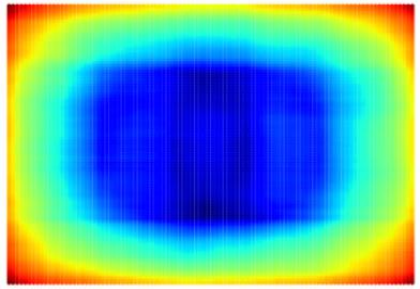
Future works

- Take advantage of PIF in **real-world defect-detection** applications.
- Employ **non-parametric models** for preference embedding (e.g., supervised trained models).

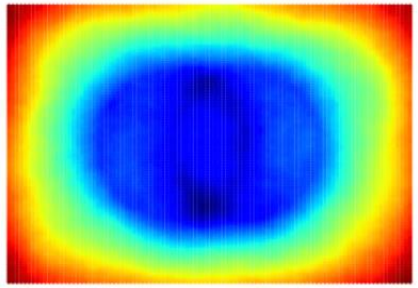
Thank you!



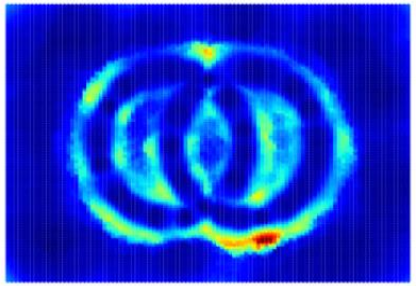
Ground truth



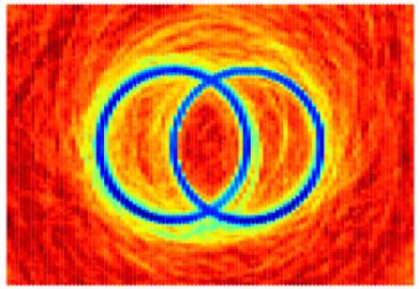
IFOR



EIFOR



LOF



PIF

If you are interested in our work, do not hesitate to drop us an e-mail!

filippo.leveni@polimi.it
luca.magri@polimi.it
giacomo.boracchi@polimi.it
cesare.alippi@polimi.it, cesare.alippi@usi.ch

POLITECNICO
MILANO 1863