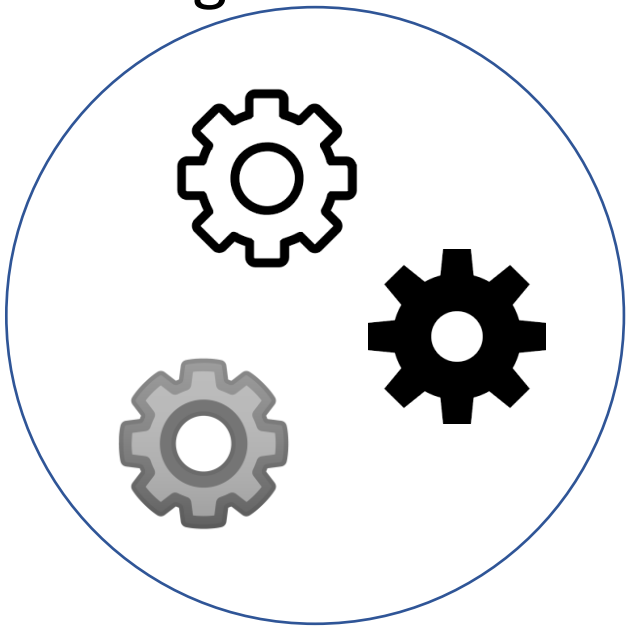
The Effect of Hyperparameter Tuning on the Comparative Evaluation of Unsupervised Anomaly Detection Methods

Jonas Soenen*, Elia Van Wolputte*, Lorenzo Perini, Vincent Vercruyssen, Wannes Meert, Jesse Davis and Hendrik Blockeel

Standard benchmarking setting

Algorithms A



Datasets D

How do algorithms A perform on datasets D?

Anomaly detection is unsupervised

How to select each algorithm's hyperparameters?

Out-of-the-box performance

Practitioners do **nothing** and use default hyperparameters

- Underestimates the potential
- Ambiguous and unfair

Tuned performance

Practitioners do **an honest effort** to select good hyperparameters

+ Realistic
+ Reproduceable, fair and sound

Peak performance

Practitioners select **optimal hyperparameters**

- Overestimates the potential
- Unsound: tuning on the test set

Tuned performance

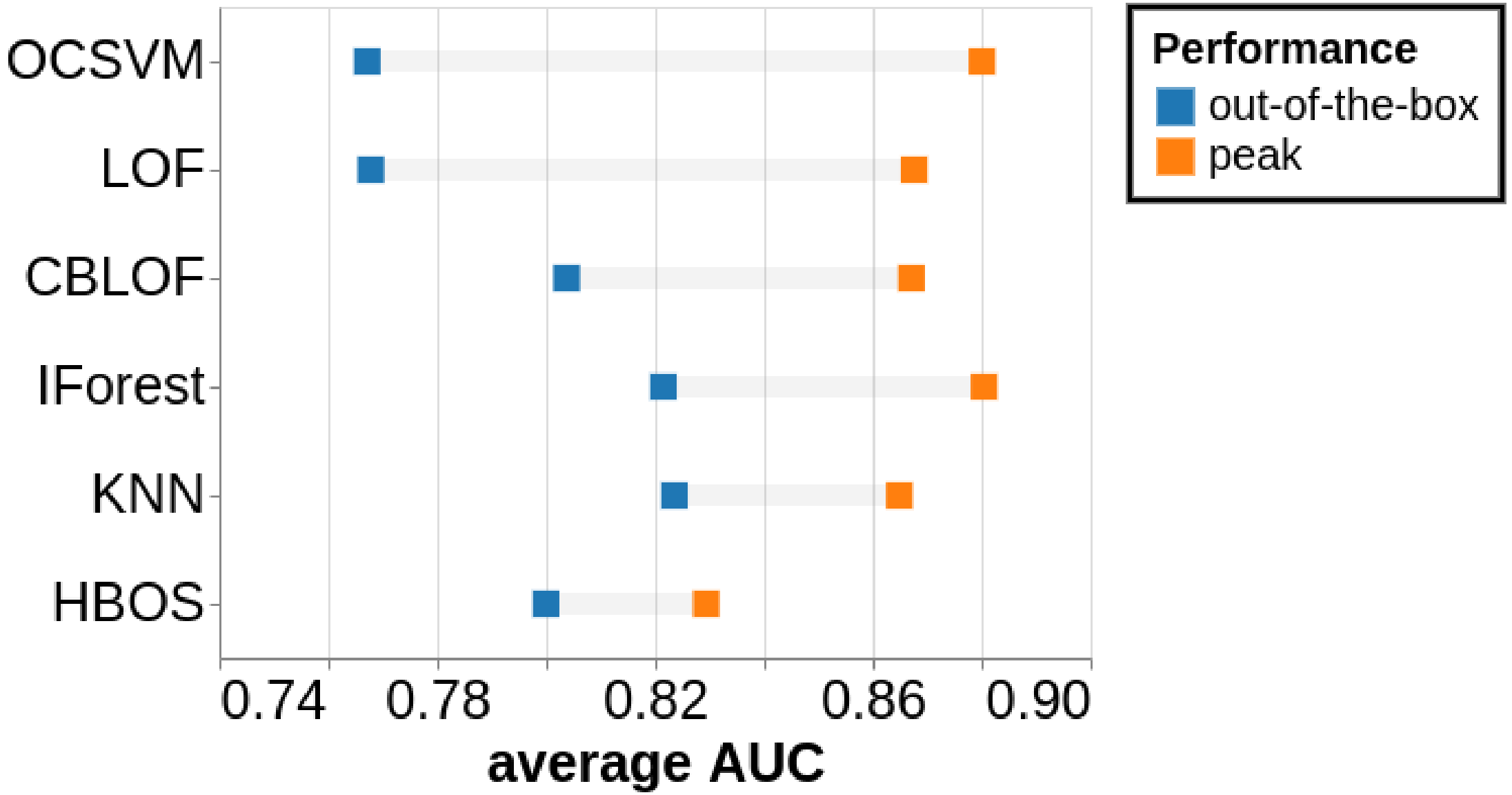
Practitioners do **an honest effort**
to select good hyperparameters

"Honest effort" to select good hyperparameters
= Tune them using a validation set

1 **Train** the anomaly detector with multiple hyperparameter settings on **unlabeled data**

2 **Select** the detector with the best performance on **a small labeled validation set**

3 **Evaluate** the selected detector based on **the labeled test set**

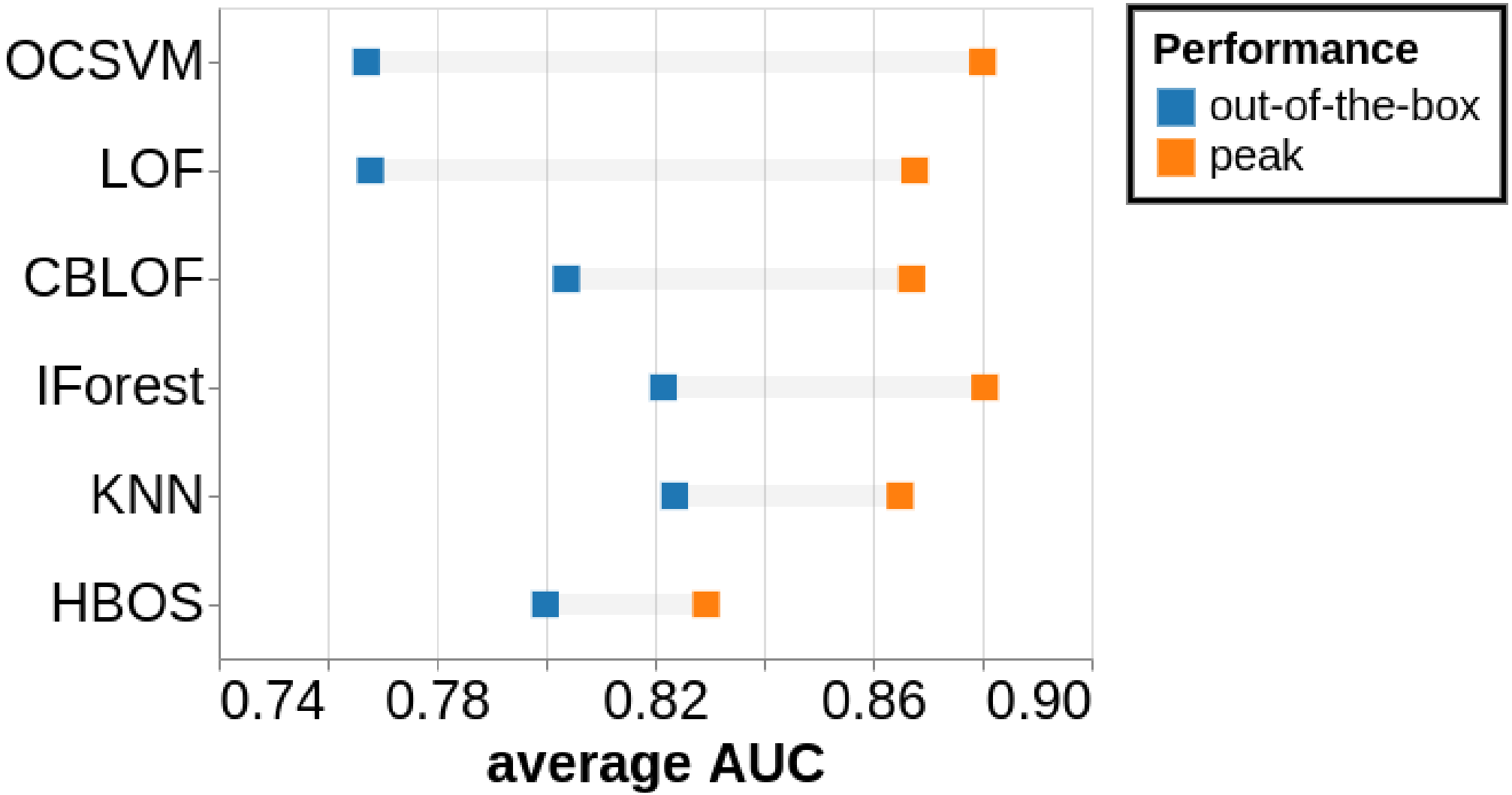


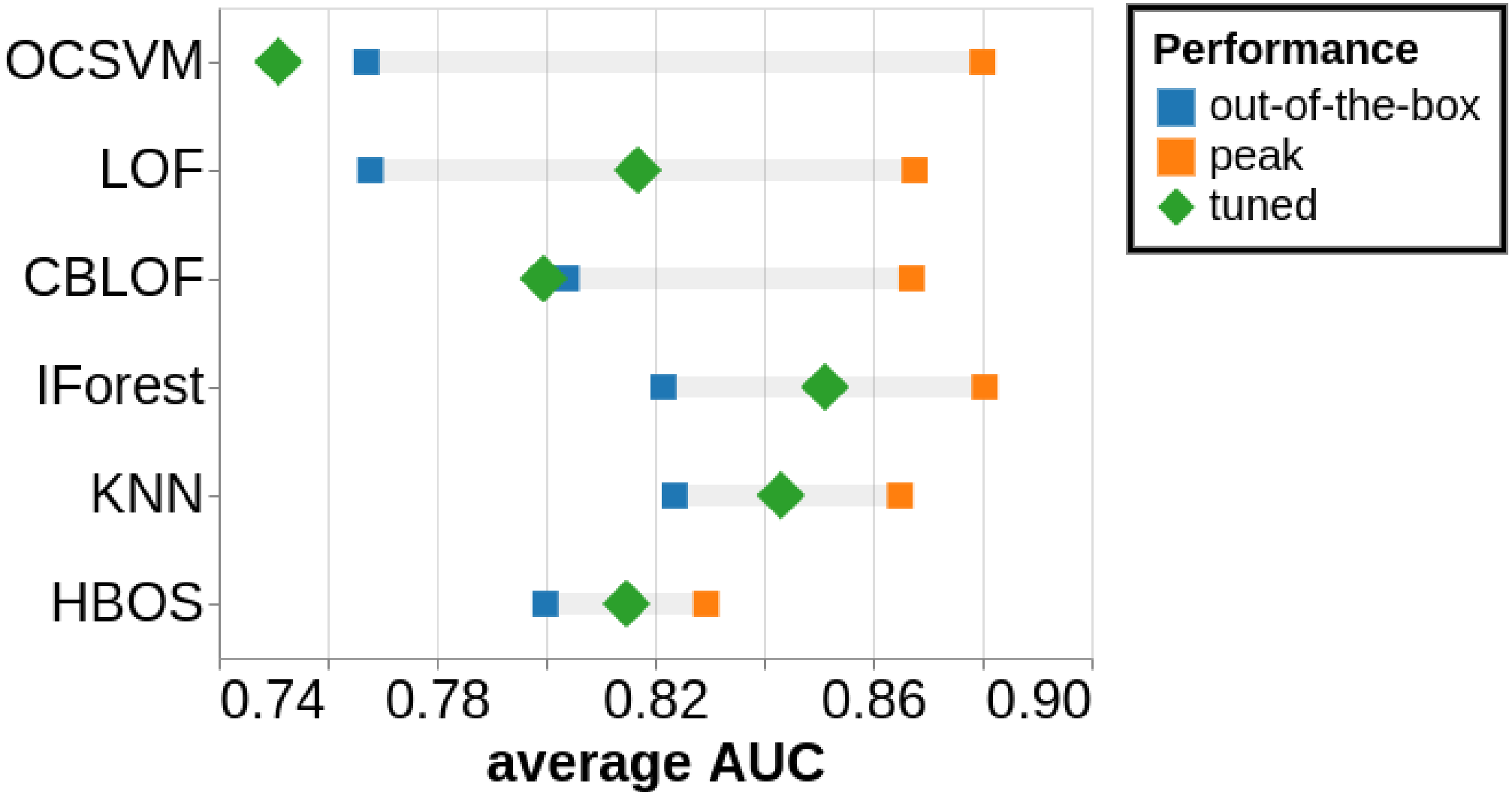
Influence of hyperparameters differs from algorithm to algorithm

Out-of-the-box (Eq. 1)		
algorithm	*avg AUC*	*rank*
IForest	0.82	2.66
KNN	0.82	2.91
CBLOF	0.8	3.22
HBOS	0.8	3.81
LOF	0.77	4.12
OCSVM	0.77	4.28

Peak (Eq. 2)		
algorithm	*avg AUC*	*rank*
IForest	0.88	2.72
CBLOF	0.87	3.03
OCSVM	0.88	3.28
LOF	0.87	3.5
KNN	0.86	3.88
HBOS	0.83	4.59

Hyperparameter selection influences algorithm ranking


OCSVM
LOF
CBLOF
IForest
KNN
HBOS
0.74
0.78
0.82
0.86
0.90
average AUC
Performance
out-of-the-box
peak



- For most algorithms, tuning helps but doesn't reach peak performance
- For others, tuning is counterproductive

Summary

Out-of-the-box performance
(default hyperparameters)

- Underestimation
- Ambiguous

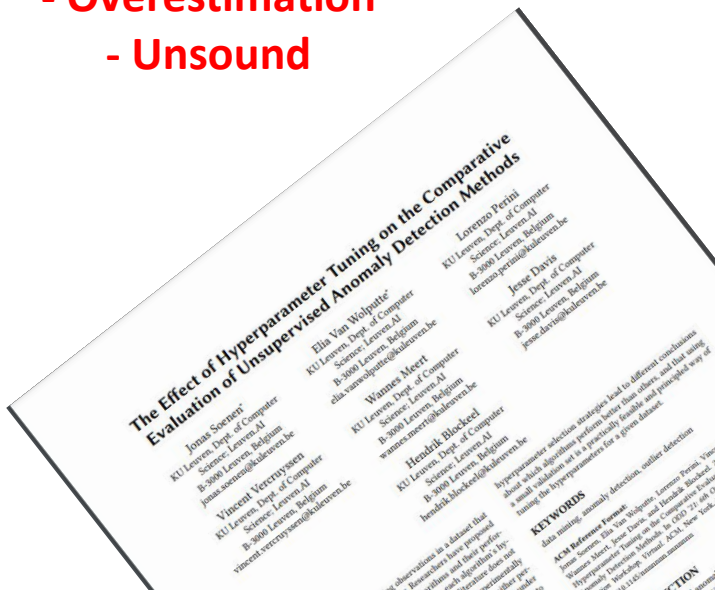
Tuned performance
(tuning on validation set)

+ Realistic
+ Reproduceable, fair and sound

Peak performance
(optimal hyperparameters)

- Overestimation
- Unsound

For details and more experiments, see the full paper


The Effect of Hyperparameter Tuning on the Comparative Evaluation of Unsupervised Anomaly Detection Methods

Jonas Soenen, Elia Van Wolputte, Lorenzo Perini, Vincent Vercruyssen, Wannes Meert, Jesse Davis, Hendrik Blockeel

KU Leuven, Dept. of Computer Science, Leuven.AI, B-3000 Leuven, Belgium

KEYWORDS