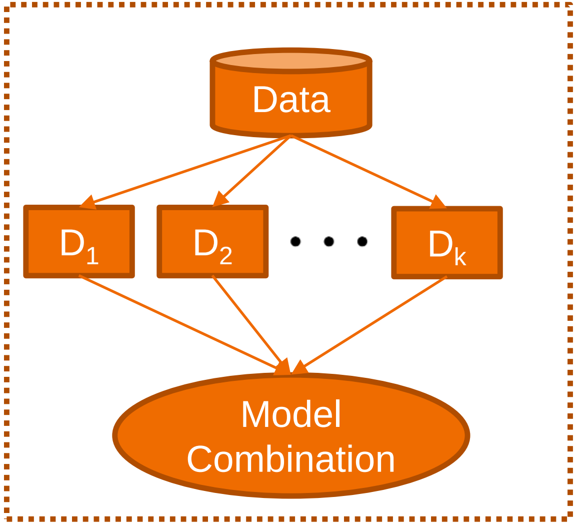
DCSO: Dynamic Combination of Detector Scores for Outlier Ensembles

Yue Zhao
Department of Computer Science
University of Toronto

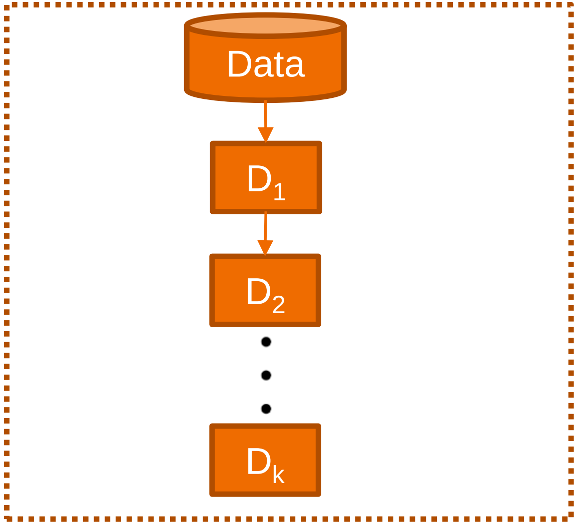
Maciej K. Hryniewicki
Data Assurance & Analytics
PricewaterhouseCoopers

Outlier Ensembles

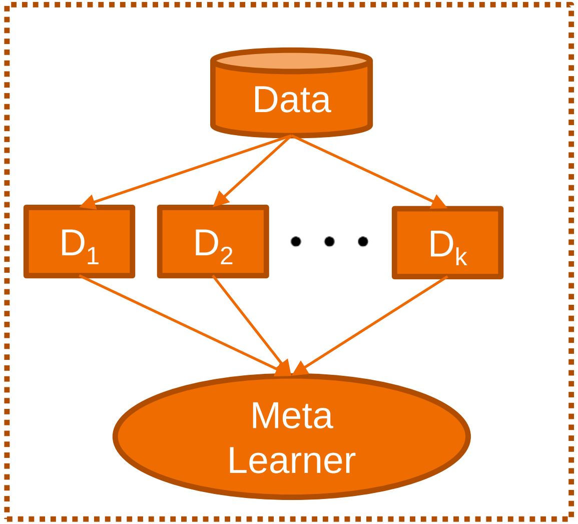
Outlier ensembles **combine** the results (scores) of either **independent** or **dependent** outlier detectors [1].



Bagging (parallel learning) [2, 3]



Boosting (sequential learning) [4, 5]



Stacking [6,7]

Advantages of Outlier Ensembles

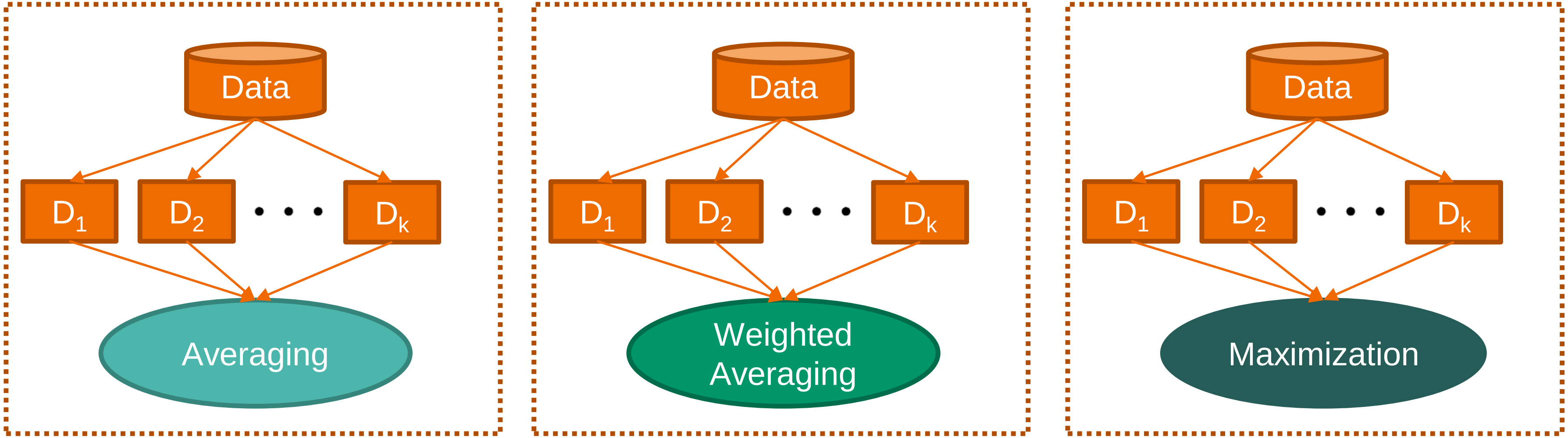
- **Improved stability:** robust to uncertainties in complex data, e.g. high-dimensional data
- **Enhanced detection quality:** capable of leveraging the advantages of underlying models

Besides, practitioners usually feel more **confident** to use an ensemble framework with a group base detectors, than a single model.

Challenges in Outlier Ensembles

The ground truth (label), whether a data object is abnormal, is always **absent**.

Most unsupervised outlier ensembles are therefore **parallel combination**.



Examples of Parallel Detector Combination

Limitations in Parallel Outlier Score Combination

- **Static process**: the process to measure detector competency is missing
- **Global assumption**: the importance of the data locality is underestimated
- **Limited interpretability**: the explicability of is undermined during combination

Static & Global Combination (**SG**): conducted **statically** on the **global** scale with all data objects considered, resulting in limited **performance** and **interpretability**.

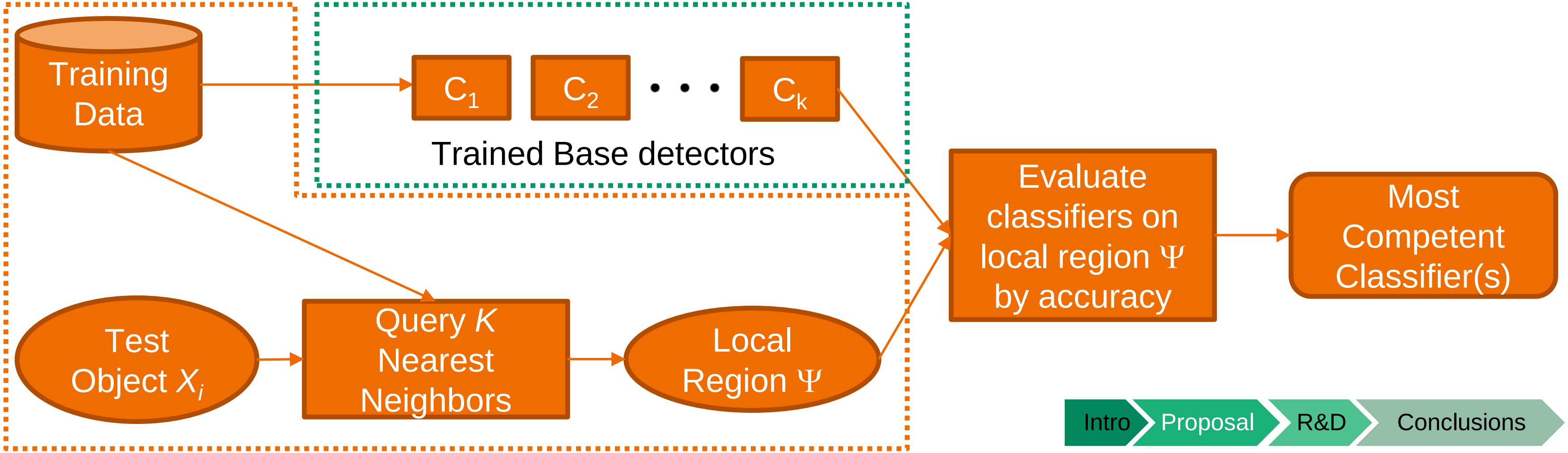
Research Objectives

Design an ***unsupervised*** combination framework to ***select performing detectors*** with a focus on ***the local region***, for improved performance and interpretability.

DCSO: **D**ynamic **C**ombination of Detector **S**cores for **O**utlier Ensembles

Dynamic Classifier Selection (DCS)

DCS is a well-established ensemble framework, which **selects the best classifier** for each test instance **on the fly** by **evaluating base classifiers' competency** on the **local region** of the test instance.

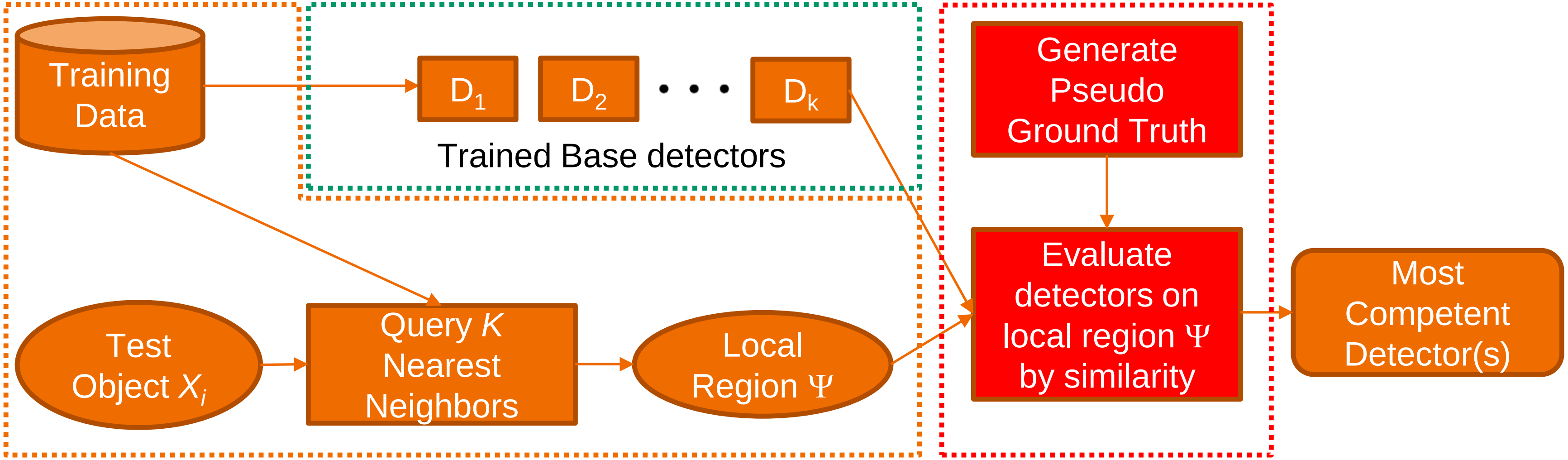


From DCS → DCSO

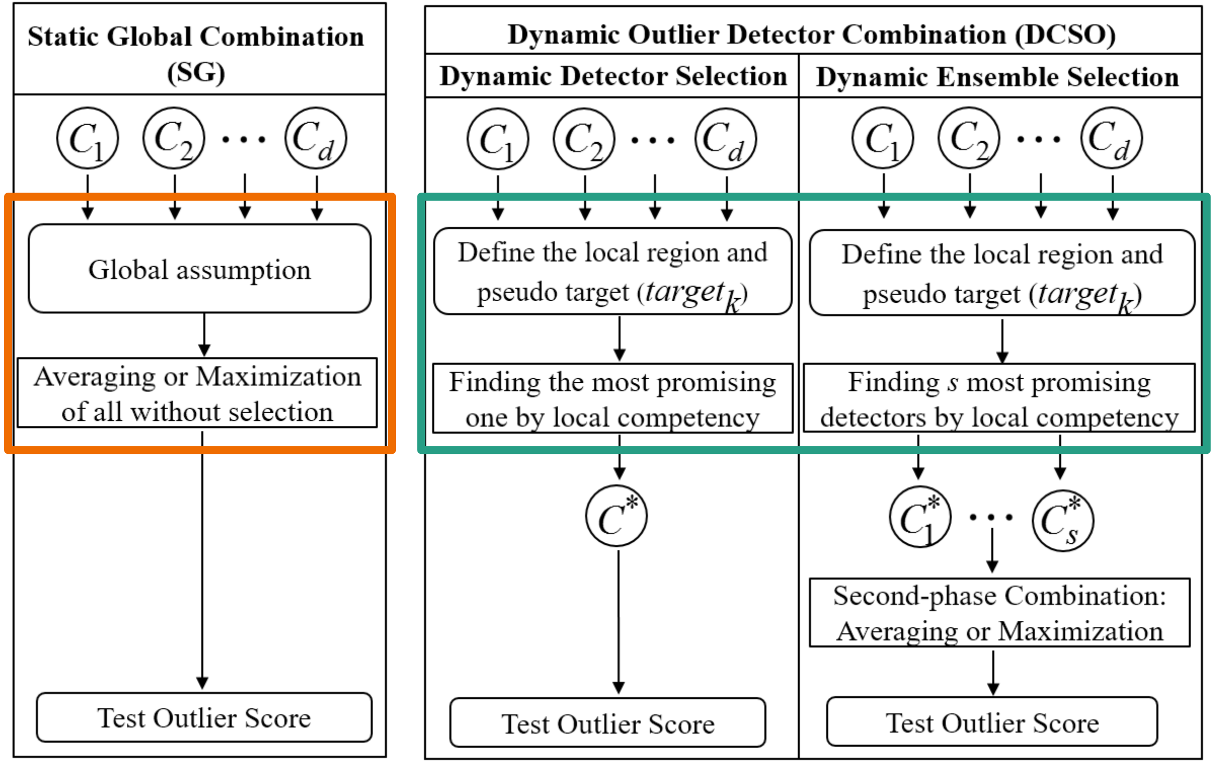
DCS (Supervised Classification)	**DCSO (Unsupervised Outlier Mining) (Imbalanced Data: outliers << inliers)**
The ground truth **exists**	The ground truth is **missing** Generate **pseudo ground truth** instead
Evaluate by **accuracy**	Evaluate the detector competency by its **similarity to the pseudo ground truth**

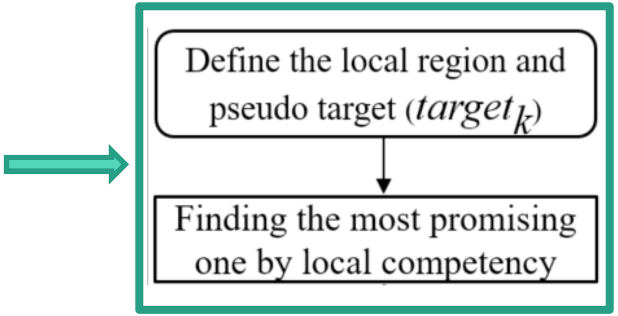
DCSO Demonstration

Different from DCS, DCSO has an additional process to generate pseudo labels, and different competency evaluation approach.



SG (left) vs. DCSO (mid) & Key Difference (right)



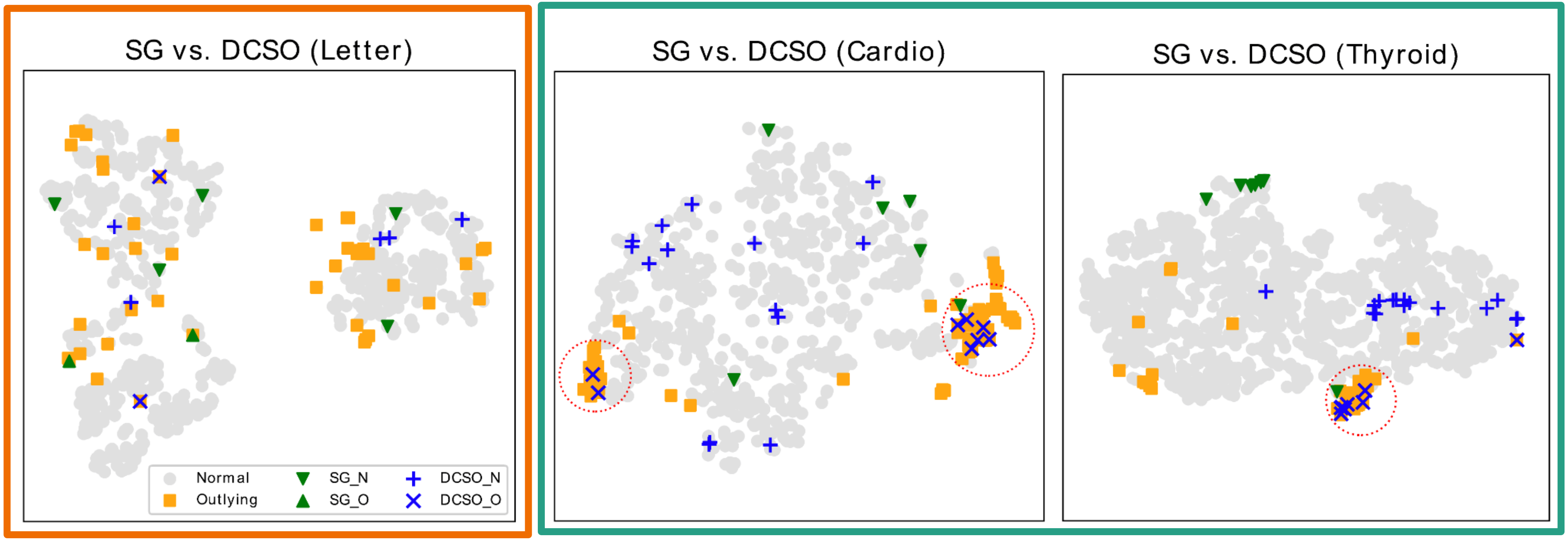


Results & Discussions – Overall Performance

Dataset	SG_A	SG_M	SG_WA	SG_ THRESH	SG_ AOM	SG_ MOA	DCSO_A	DCSO_M	DCSO_ MOA	DCSO_ AOM
Pima	0.5100	0.4683	0.5127	0.4933	0.4957	0.5039	**0.5175**	0.4576	0.5083	0.4576*
Vowels	0.3074	0.3250	0.3029*	0.3074	0.3302	0.3185	**0.3682**	0.3044	0.3395	0.3161
Letter	0.2508	0.3547	0.2469	0.2508	0.2950	0.2699	0.2426*	**0.3795**	0.2862	0.3785
Cardio	0.3601	0.3733	0.3624	0.3728	0.4233	0.4104	0.3553	0.3676	**0.4453**	0.3201*
Thyroid	0.3936	0.2589	0.4061	0.3968	0.3731	0.3896	**0.4182**	0.2080*	0.3730	0.2449
Satellite	0.4301*	0.4500	0.4306	0.4466	0.4480	0.4414	0.4400	0.4427	**0.4509**	0.4398
Pendigits	0.0733	0.0590	0.0709	0.0700	0.0637	0.0617	0.0749	0.0595	**0.0811**	0.0560*
Annthyroid	0.2943	0.2951	0.2975	0.2997	**0.3215**	0.3103	0.3065	0.2904*	0.3075	0.3046
Mnist	0.3936	0.3737	0.3944	0.3956	0.3966	0.3976	0.3973	0.3541	**0.4123**	0.3520*
Shuttle	0.1508	0.1484	0.1434	0.1582	0.1591	0.1600	0.1589	0.1389*	**0.1604**	0.1393

- DCSO frameworks outperform on 8 out of 10 datasets for both *ROC* and *Precision @ Rank N*
- In generally, DCSO brings consistent improvement over baselines, and significant enhancement on *Cardio* (25.33%) and *Pendigits* (31.44%).

Results & Discussions – When DCSO Works?



Visualization by t-distributed stochastic neighbor embedding (TSNE)

DCSO works especially well when data forms local clusters.

Conclusion

DCSO is an ensemble framework to **select outperforming base detectors** for each test instance on **its local region**.

Advantages:

1. Outperform on most of the benchmark datasets with improved detection quality
2. Easy to use and robust to underlying assumptions
3. Better interpretability to show how the prediction is made individually

Code & Outlier Detection Toolbox

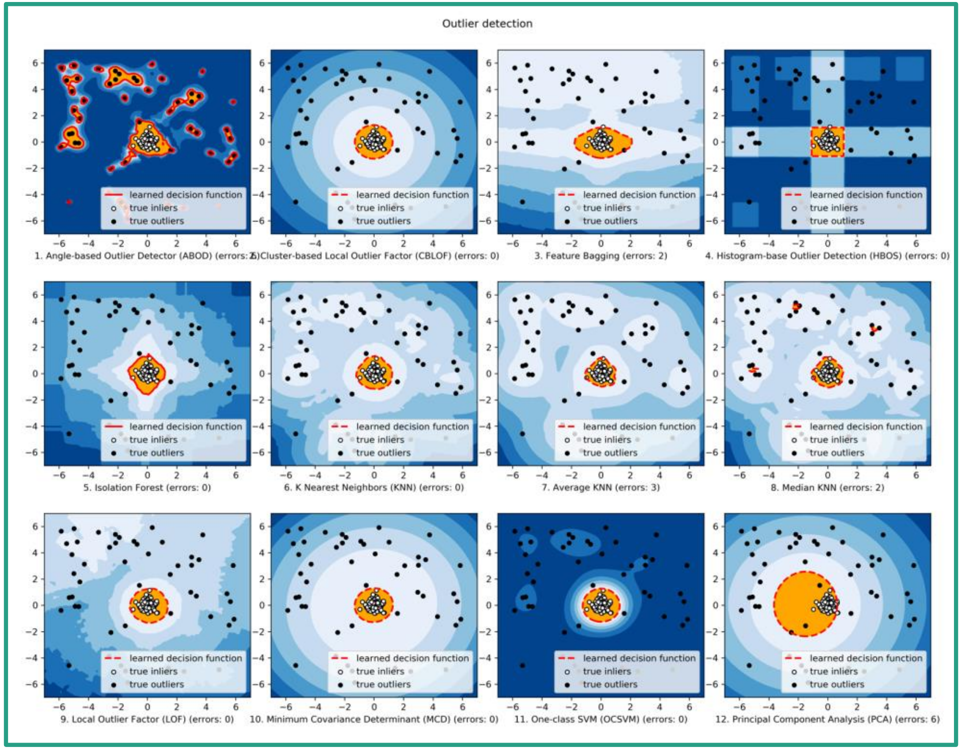
DCSO code is openly shared at:

https://github.com/yzhao062/DCSO

PyOD: a comprehensive Python outlier detection toolbox:

https://github.com/yzhao062/Pyod

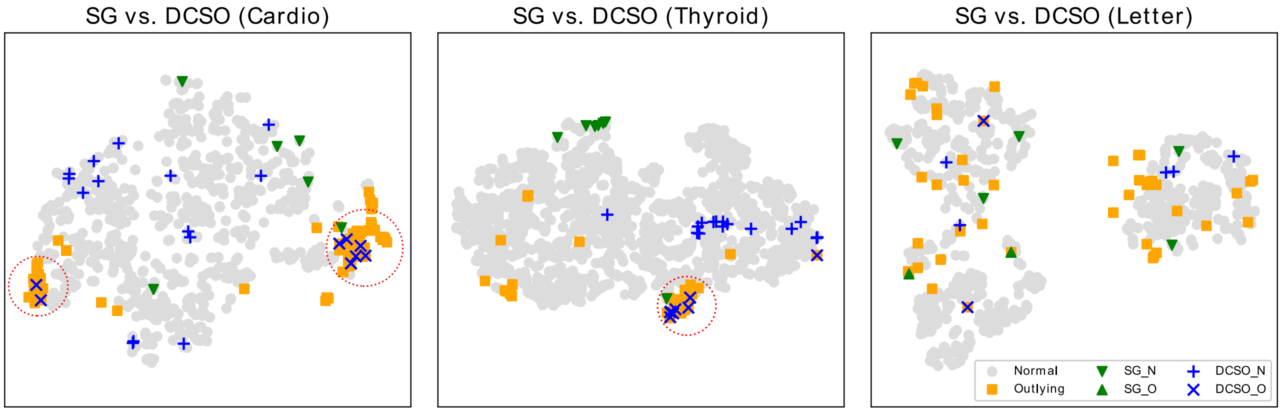
Google: “Python” + “Outlier Detection”



DCSO: Dynamic Combination of Detector Scores for Outlier Ensembles

Yue Zhao
Department of Computer Science
University of Toronto

Maciej K. Hryniewicki
Data Assurance & Analytics
PricewaterhouseCoopers



UNIVERSITY OF TORONTO

Limitations & Future Directions

Limitation: high time complexity using kNN for defining the local region

Future direction: define the local region by clustering instead

Limitation: pseudo generation methods are not accurate

Future direction: involve more advanced generation methods

Limitation: focusing on homogeneous base detectors only

Future direction: include heterogeneous base detectors for more diversity

Reference

[1] Aggarwal, C.C. 2013. Outlier ensembles: position paper. *ACM SIGKDD Explorations*. 14, 2 (2013), 49–58.

[2] Lazarevic, A. and Kumar, V. 2005. Feature bagging for outlier detection. *ACM SIGKDD*. (2005), 157.

[3] Liu, F.T., Ting, K.M. and Zhou, Z.H. 2008. Isolation forest. *ICDM*. (2008), 413–422.

[4] Rayana, S. and Akoglu, L. 2016. Less is More: Building Selective Anomaly Ensembles. *TKDD*. 10, 4 (2016), 1–33.

[5] Rayana, S., Zhong, W. and Akoglu, L. 2017. Sequential ensemble learning for outlier detection: A bias-variance perspective. *ICDM*. (2017), 1167–1172.

[6] Micenková, B., McWilliams, B. and Assent, I. 2015. Learning Representations for Outlier Detection on a Budget. arXiv Preprint arXiv:1507.08104.

[7] Zhao, Y. and Hryniewicki, M.K. 2018. XGBOD: Improving Supervised Outlier Detection with Unsupervised Representation Learning. *IJCNN*. (2018).