Informed Novelty Detection in Sequential Data by Per-Cluster Modeling

Linara Adilova [1]
linara.adilova@rub.de

Siming Chen [2]
simingchen@fudan.edu.cn

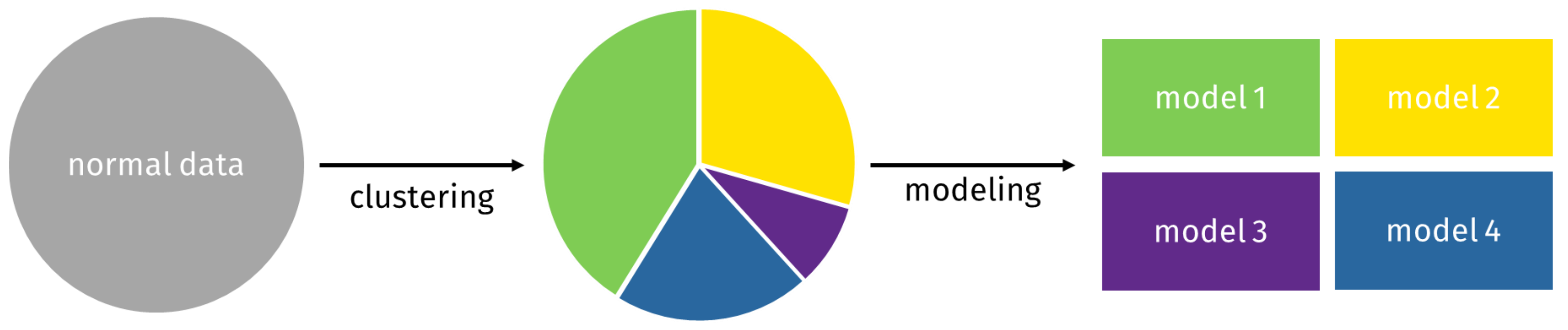
Michael Kamp [1,3,4]
michael.kamp@uk-essen.de

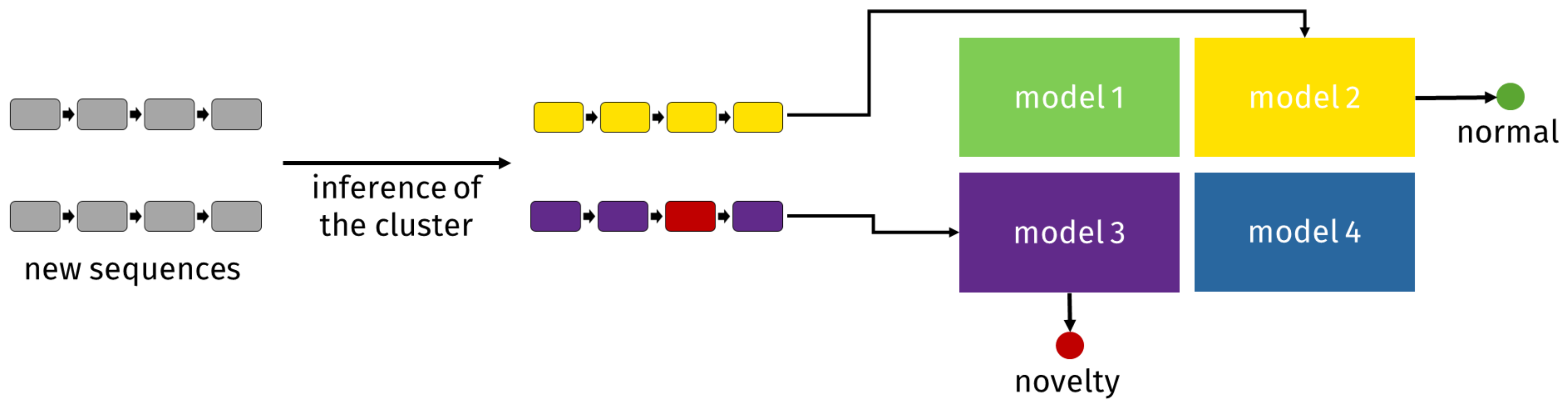
[1] Ruhr University Bochum, Germany
[2] Fudan University, China
[3] Monash University, Australia
[4] Institute for AI in Medicine (IKIM), Germany

Cluster first, then model data!

Human interaction outperforms automatic clustering: informed clustering by experts via visual analytics.

Novelty detection in discrete sequences





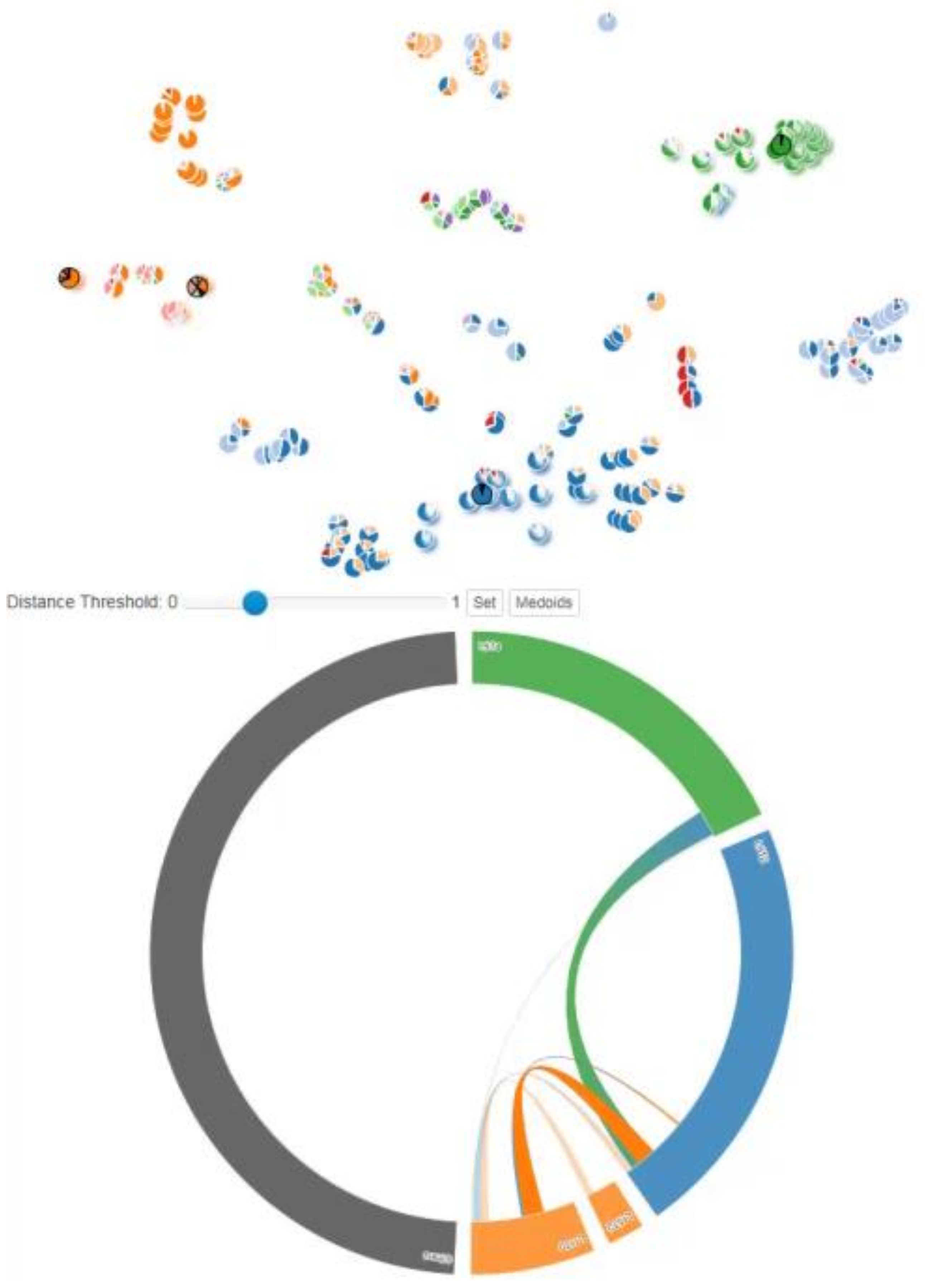
Algorithm 1 Novelty Detection via Per-Cluster Modelling

Input dataset $X \subset \mathcal{V}^*$, threshold $\theta \in \mathbb{R}$, sequence $s' \in \mathcal{V}^*$

Output $\{0, 1\}$ (0 for a normal sequence, 1 for a novelty)

```
Training:
obtain clustering $C$ with $k$ clusters of $X$
for $\mathcal{G}_i = \{s \in X \mid C(s) = i\}$ with $i = 1, \ldots, k$ do
    train process model $h_i$ on $\mathcal{G}_i$
end for
Inference:
compute cluster $C(s')$ of $s'$
if $PP(h_{C(s')}, s') > \theta$ then
    return 1
else
    return 0
end if
```

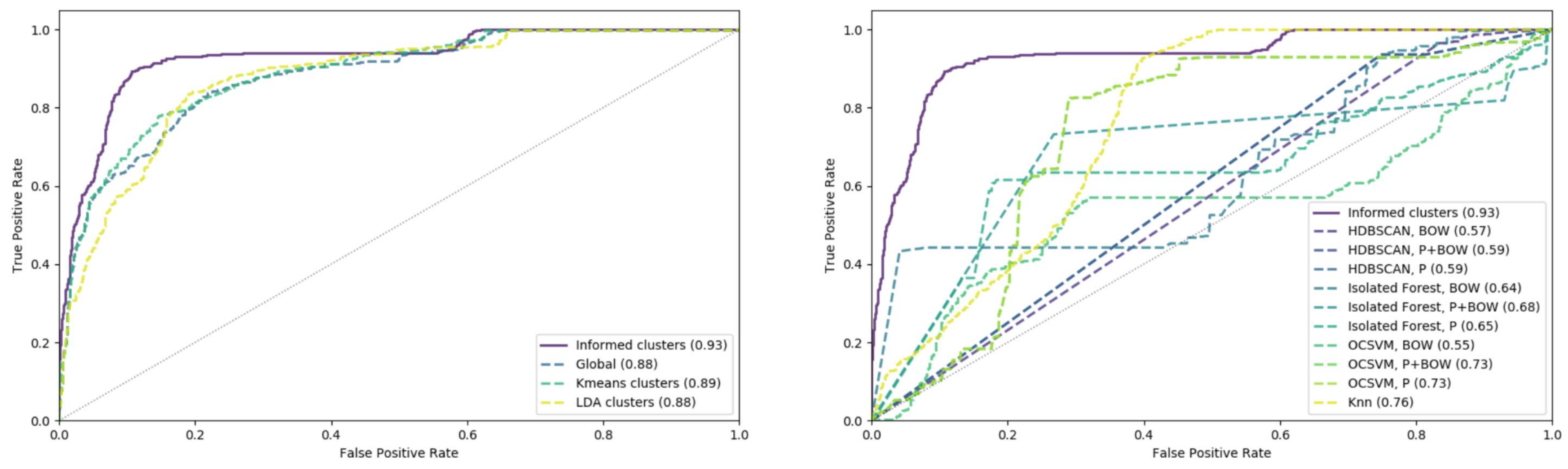
Expert-informed clustering



For a detailed demonstration see

Evaluation

Cybersecurity dataset



Method	**AUC**	$\frac{\textbf{Sens.+Spec.}}{\textbf{2}}$	Sens.	Spec.
IC-LSTMs	**0.93**	**0.89**	0.89	0.89
Global LSTM	0.88	0.81	0.80	0.81
k-means Cluster LSTMs	0.89	0.81	0.83	0.78
LDA Cluster LSTMs	0.88	0.82	0.81	0.82

Fake reviews dataset

Method	**AUC**	$\frac{\textbf{Sens.+Spec.}}{\textbf{2}}$	Sens.	Spec.
IC-LSTMs	**0.58**	**0.58**	0.58	0.57
Global LSTM	0.55	0.54	0.53	0.55
k-means Cluster LSTMs	0.52	0.51	0.50	0.52
LDA Cluster LSTMs	**0.58**	0.57	0.57	0.57

CPU utilization timeseries

Method	**AUC**	$\frac{\textbf{Sens.+Spec.}}{\textbf{2}}$	Sens.	Spec.
IC-LSTMs	**0.99**	**0.87**	0.77	0.97
Global LSTM	0.96	0.86	0.77	0.94
k-means Cluster LSTMs	0.97	0.84	0.85	0.82
LDA Cluster LSTMs	0.98	**0.87**	0.77	0.97