xSTREAM | Outlier Detexion in Feature-Evolving Data STREAMs


Emaad Manzoor

Hemank Lamba

Leman Akoglu

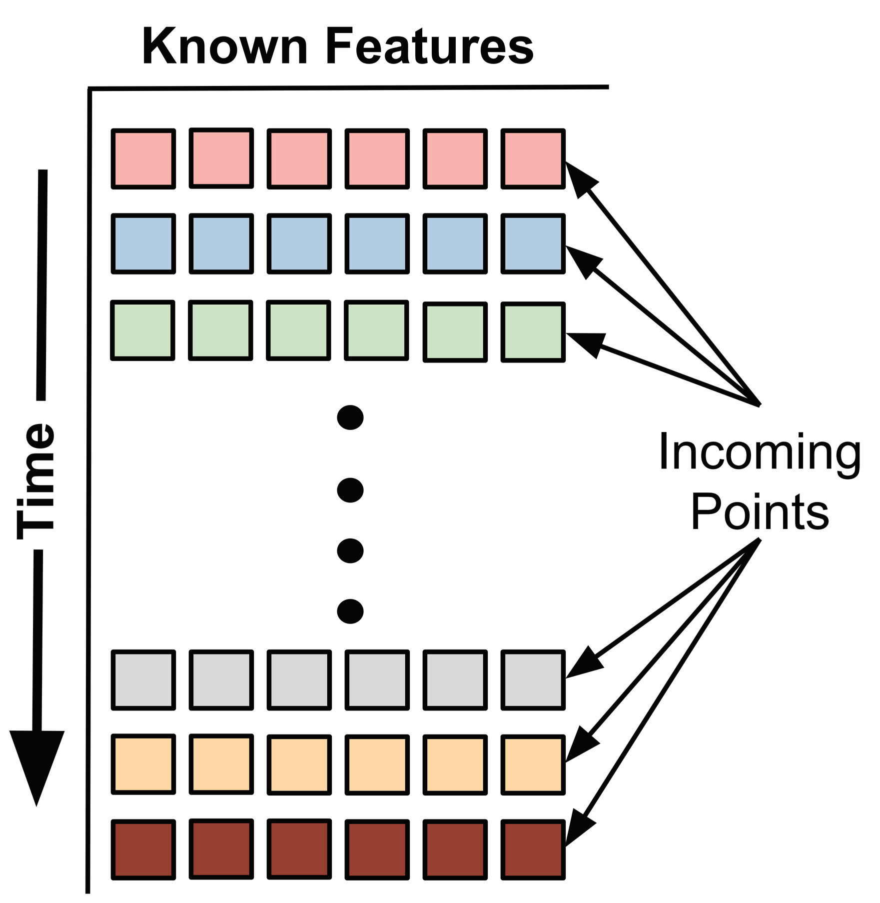
Code and data available at **cmuxstream.github.io**

Carnegie Mellon University

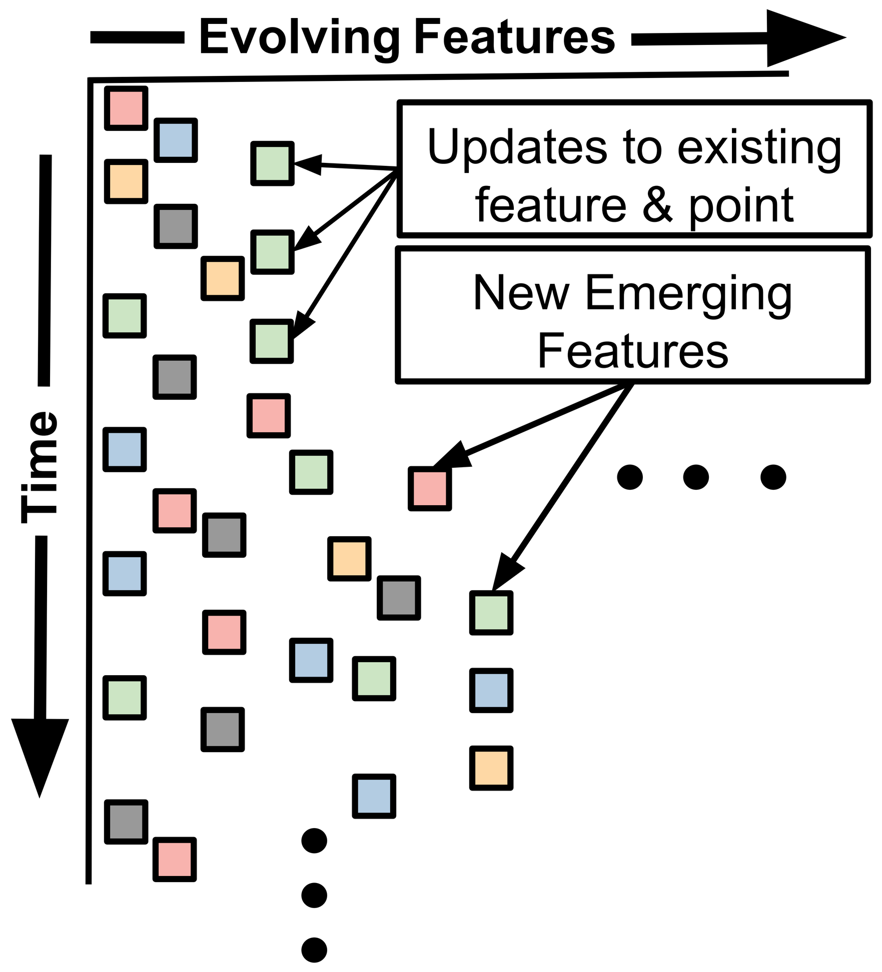
xStream detects outliers in dynamic streams having a large and evolving feature-space

#kdd2018 #icml2018 #nips2018

#kdd2018 will feature keynotes by David Hand, Alvin Roth, @yeewhye and Jeannette Wing … | Towards Actionable Intelligence - this **#icml2018** tutorial was so good! … | … do you like the acceptance in **#nips2018** to be decided by random noise?



PREVIOUS METHODS



THIS WORK: XSTREAM

Challenges

Large and evolving feature-space

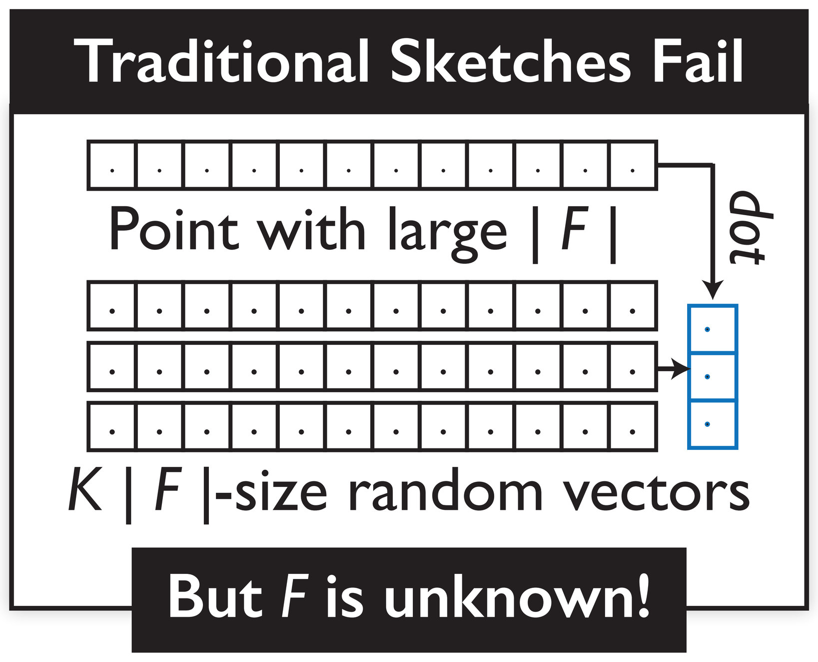
Point **updates** & **concept drift**

Outliers at **multiple granularities**

Limited memory

PROPERTIES	STORM	HSTREES	LODA	RS-HASH	RS-FOREST	XSTREAM
MULTI-SCALE						✓
SUBSPACES		✓	✓	✓	✓	✓
PROJECTIONS			✓			✓
EVOLVING POINTS						✓
EVOLVING FEATURES						✓

STREAMHASH: Sparse Streaming Sketches



Idea: don't cache, hash!

$h_i(f): f \rightarrow \{+1, 0, -1\}$

$h_1 \dots h_K$ take **constant space!**

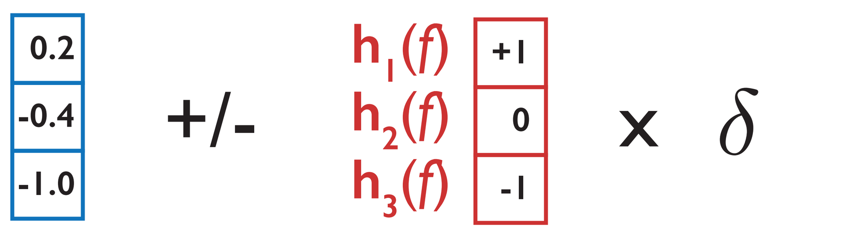
Random Subspace Selection

$$h_i[f] = \sqrt{\frac{3}{K}} \begin{cases} -1 & \text{with prob. } 1/6 \\ 0 & \text{with prob. } 2/3 \\ +1 & \text{with prob. } 1/6 \end{cases}$$

2/3 chance of feature being dropped

Constant-time Point Updates

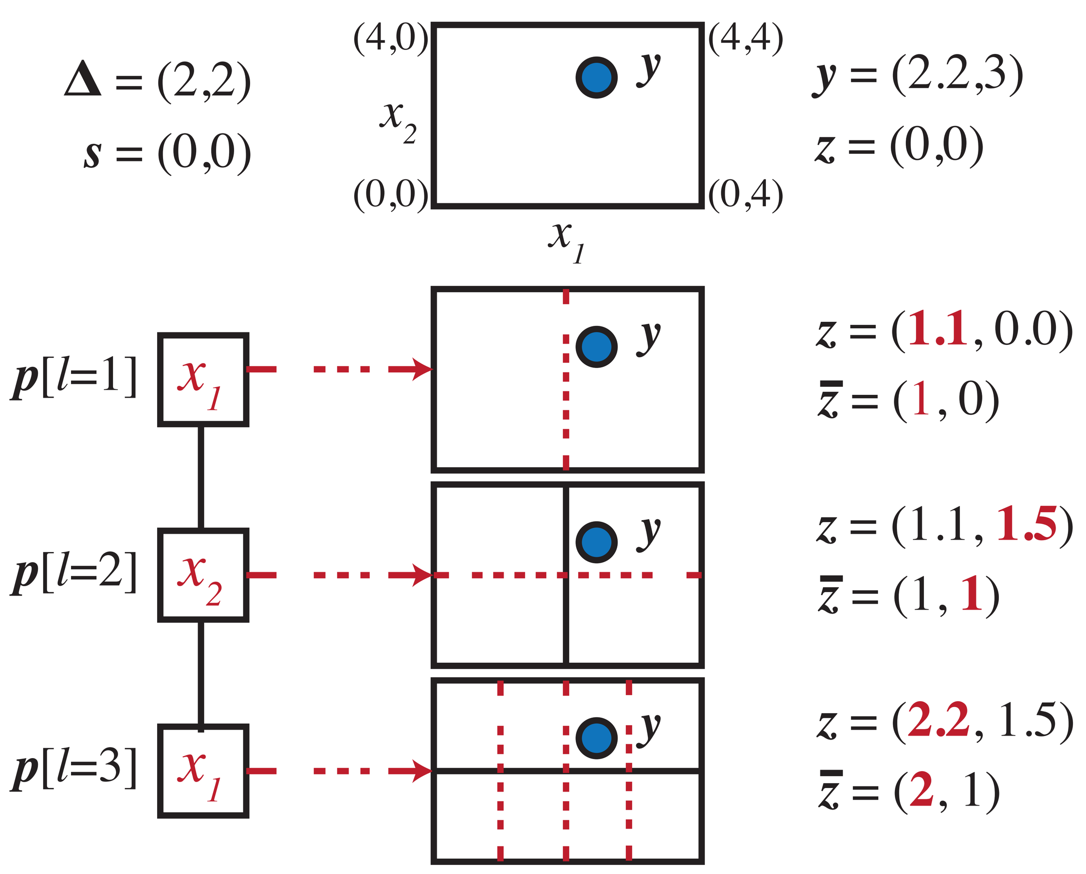
Stream update: (id, f, δ)



Projection of point *id*

Hash updates of feature *f*

Half-Space Chains



Score of each chain over all levels l

$$\text{score}(\mathbf{y}) = \min_l \; 2^l \times \text{count}_l[\bar{\mathbf{z}}]$$

Method Highlights

Density-estimation ensemble to detect outliers at **multiple scales**

Projected subspace method to detect outliers in **high + unknown dimensionality**

Alternating windows to handle **non-stationarity & concept drift**

Constant time and space complexity to handle big, rapid data streams

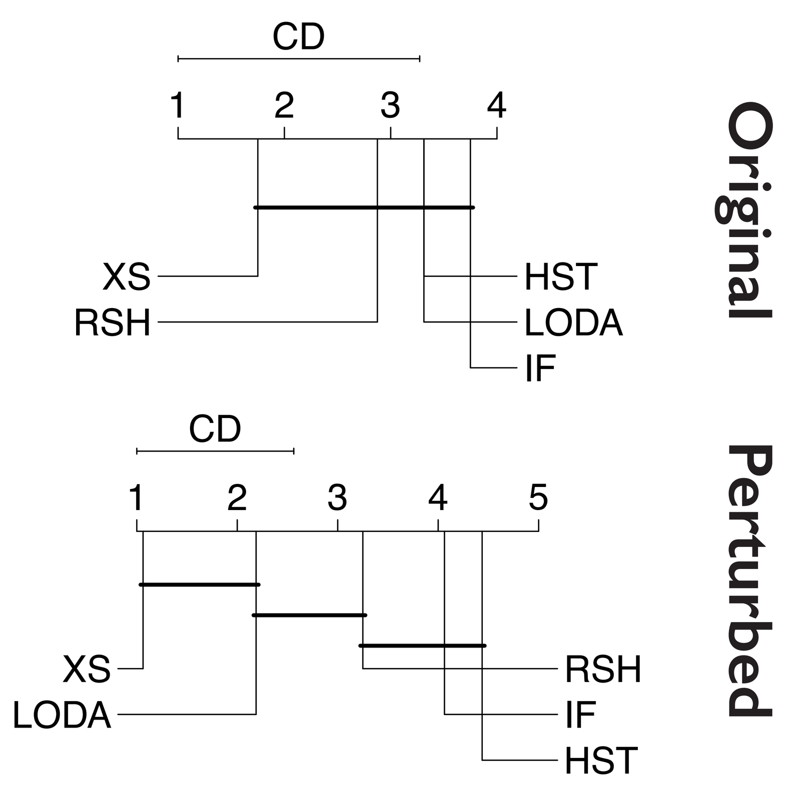
Time $O(KmDM)$

Space $O(MmLD + NK)$

Static Data

8 UCI Outlier Detection Datasets

Avg. Rank	*Original*	*Perturbed*
XSTREAM	1.75	1.06
IFOREST	3.75	4.06
HSTREES	3.31	4.44
RSHASH	2.88	3.25
LODA	3.31	2.19

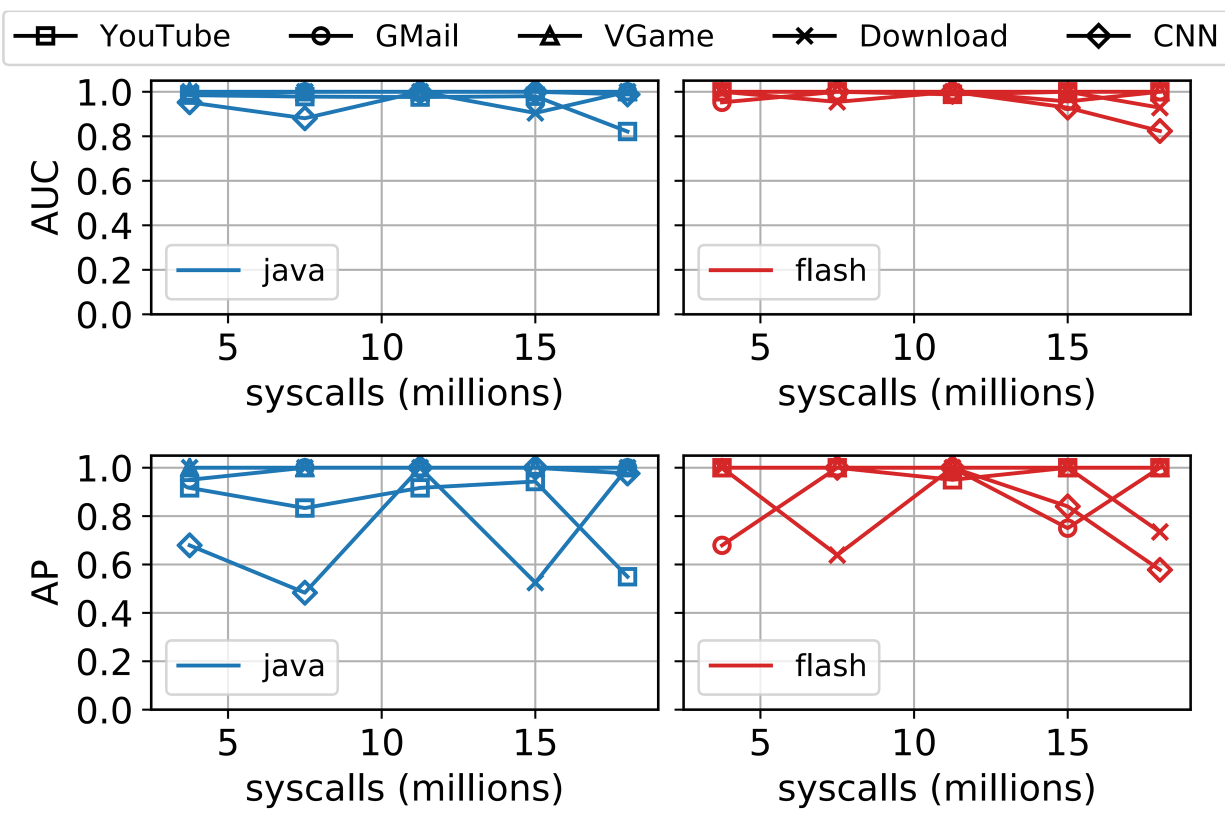


Row-stream Data

Dataset	|D|	d	# outliers
SPAM-SMS	5.5K	8.4K	747
SPAM-URL	2.4M	3.2M	792K

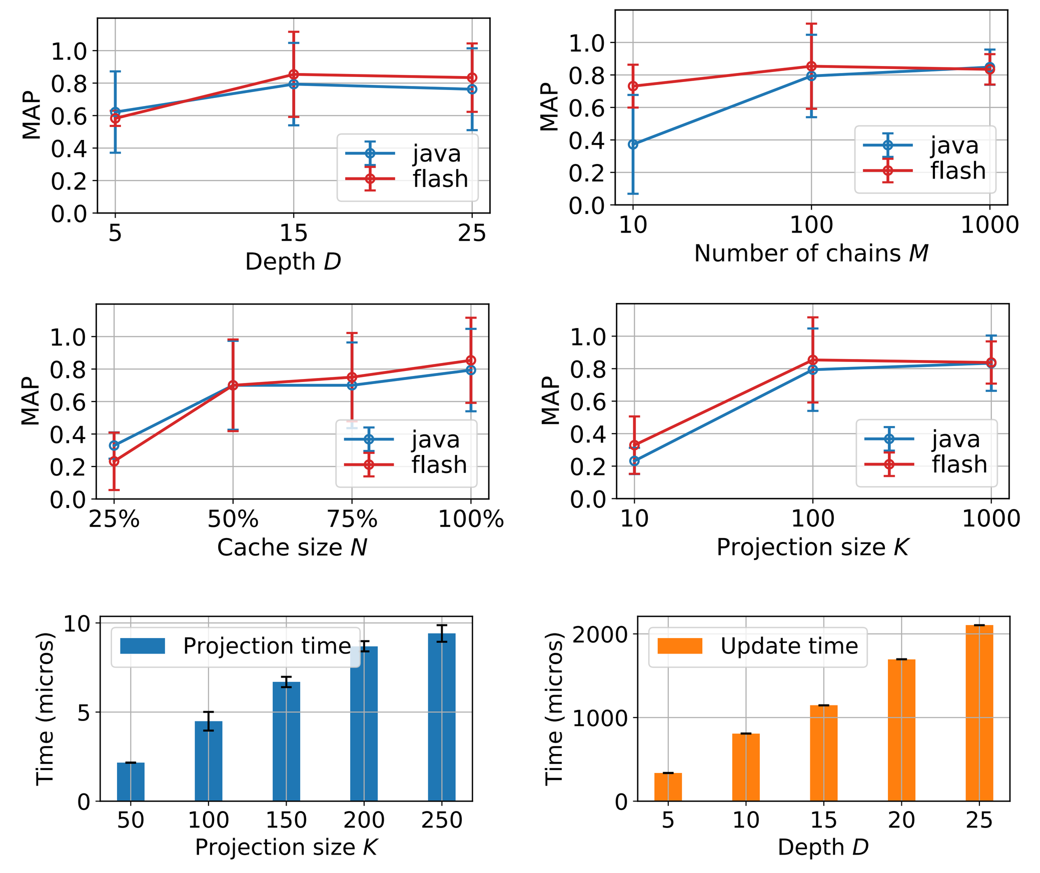
Avg. Precision	*Mean*	*Overall*
XSTREAM	**0.409**	**0.404**
HSTREES	0.363	0.359
RSHASH	0.203	0.201
LODA	0.080	0.080

Evolving-stream Data



Mean Avg. Precision	*Attk-Java*	*Attk-Flash*
All Scenarios	0.794	0.854

Dataset	*n*	*d*	*# outliers*
ATTK-FLASH	63.1M	1.1M	2.8M
ATTK-JAVA	89.7M	1.1M	29.5M



Research supported by: