UNIVERSITÄT
BIELEFELD
Technische Fakultät

Detecting and explaining drift

Barbara Hammer

Machine Learning Group, Bielefeld University

Incremental Learning Workshop at ICDM’22

UNIVERSITÄT BIELEFELD

Technische Fakultät

1) Supervised scenario: learning with streaming data and possible drift

UNIVERSITÄT BIELEFELD
Technische Fakultät

Supervised learning on data streams

Given a **stream of training data**

$(x^1, y^1), \ldots, (x^t, y^t), \ldots \in X \times Y$

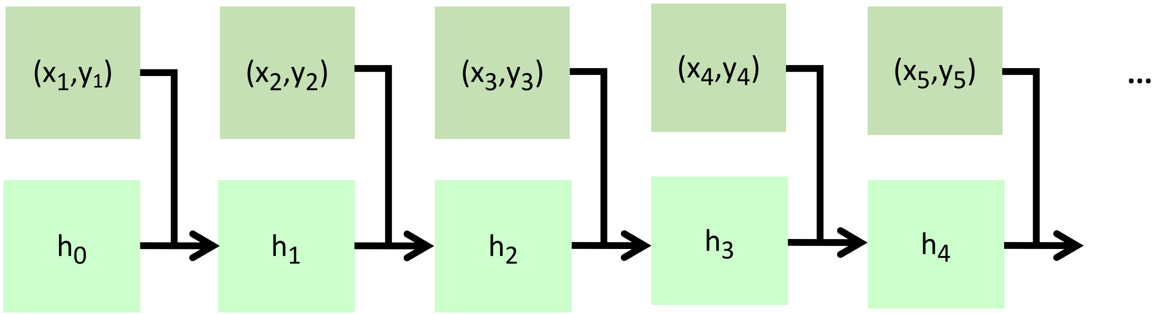
sampled w.r.t. a family of probability distributions P_t on $X \times Y$

We aim for a **learning scheme which incrementally adapts a model** $h_t : X \to Y$ based on (x^t, y^t) such that the interleaved train-test error

$E \quad = \sum_t l(h_{t-1}(x_t), y_t)$ is minimized.

UNIVERSITÄT BIELEFELD
Technische Fakultät

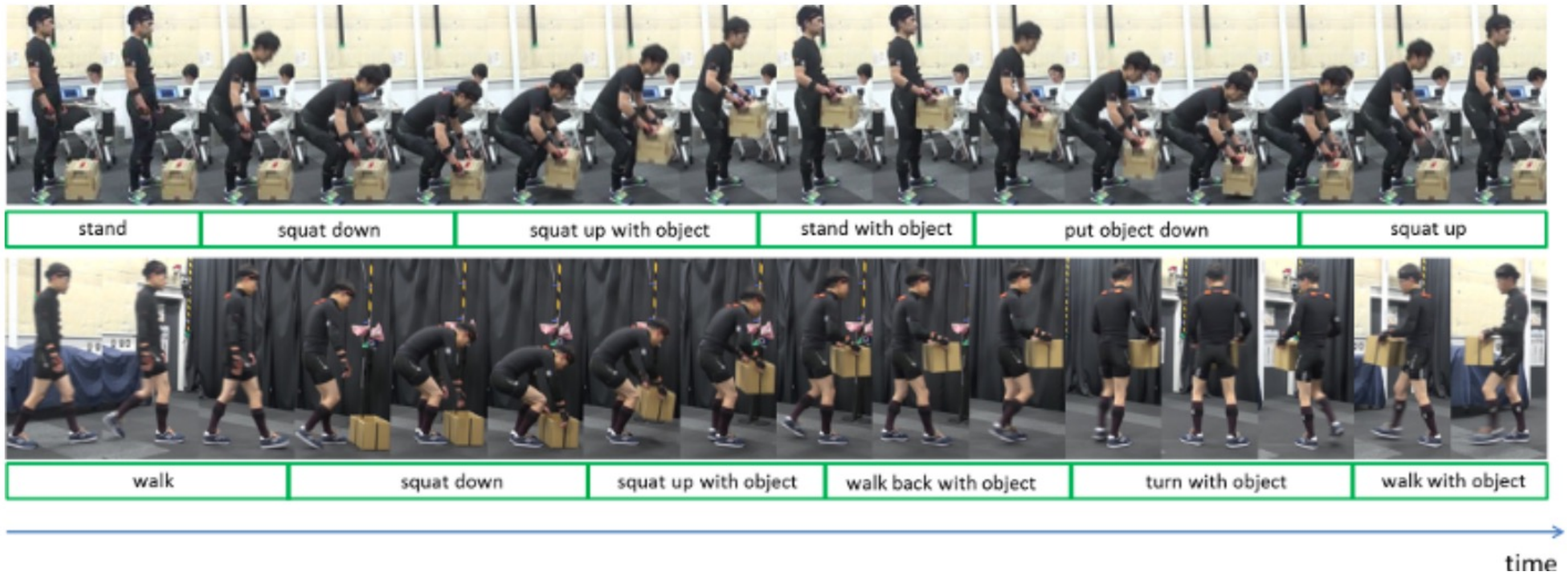
Learning from data streams



UNIVERSITÄT
BIELEFELD

Technische Fakultät

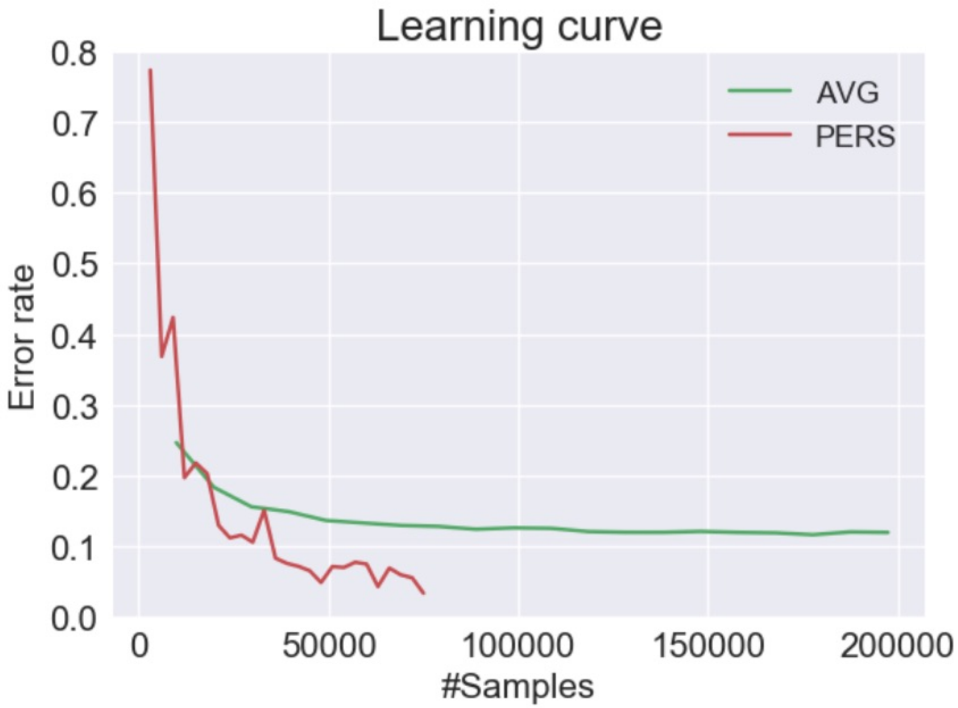
Personalized prognosis of motions



Viktor Losing, Taizo Yoshikawa, Martina Hasenjäger, Barbara Hammer, Heiko Wersing:
Personalized Online Learning of Whole-Body Motion Classes using Multiple Inertial Measurement Units. ICRA 2019: 9530-9536

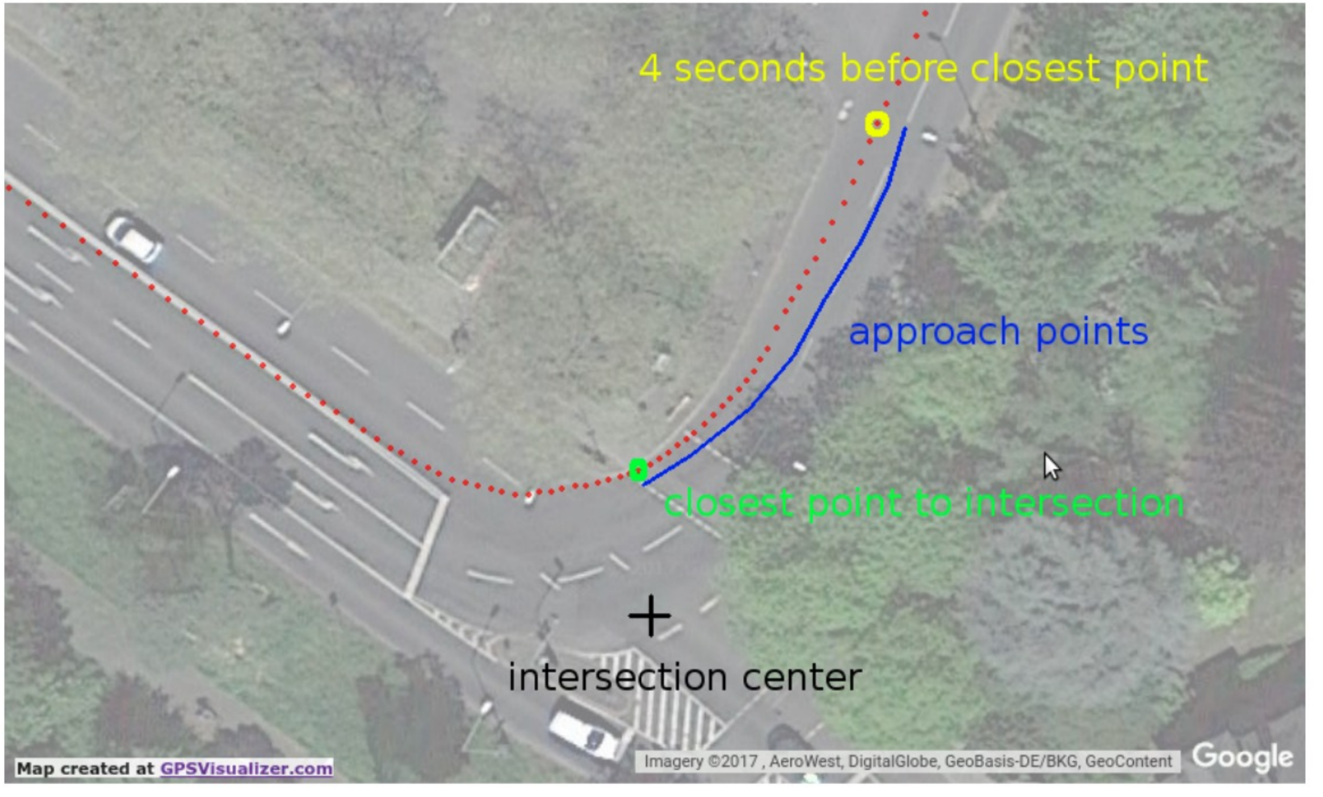
UNIVERSITÄT
BIELEFELD
Technische Fakultät

Personalized prognosis of motions



Viktor Losing, Taizo Yoshikawa, Martina Hasenjäger, Barbara Hammer, Heiko Wersing:
Personalized Online Learning of Whole-Body Motion Classes using Multiple Inertial Measurement Units. ICRA 2019: 9530-9536

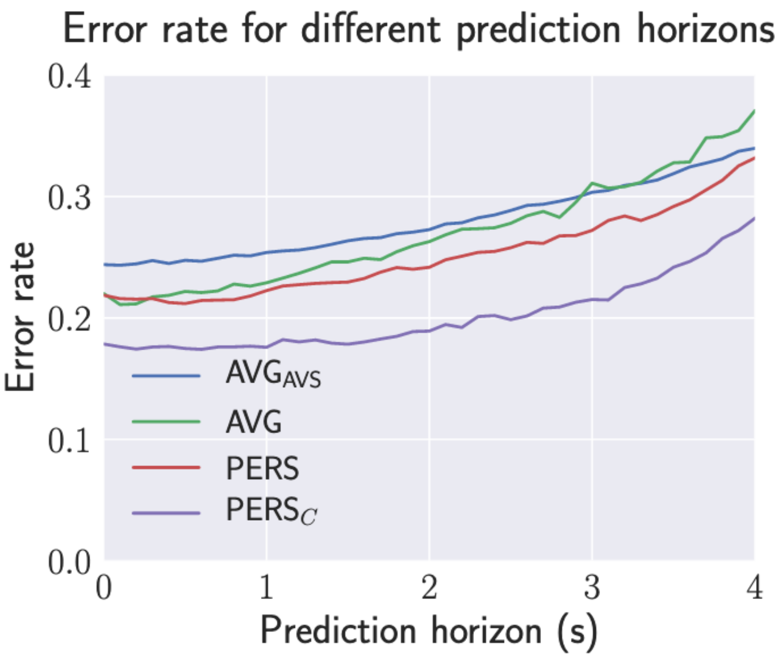
Personalized assistant for crossings

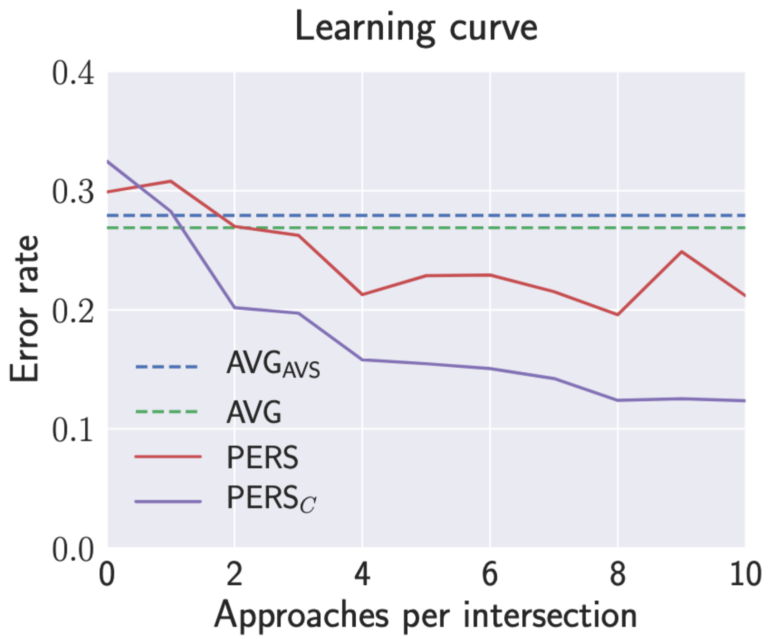


Losing, Hammer, Wersing
"Personalized Maneuver Prediction
at Intersections", ITSC 2017

UNIVERSITÄT
BIELEFELD
Technische Fakultät

Personalized assistant for crossings





Losing, Hammer, Wersing
"Personalized Maneuver Prediction
at Intersections", ITSC 2017

UNIVERSITÄT BIELEFELD
Technische Fakultät

Supervised learning on data streams

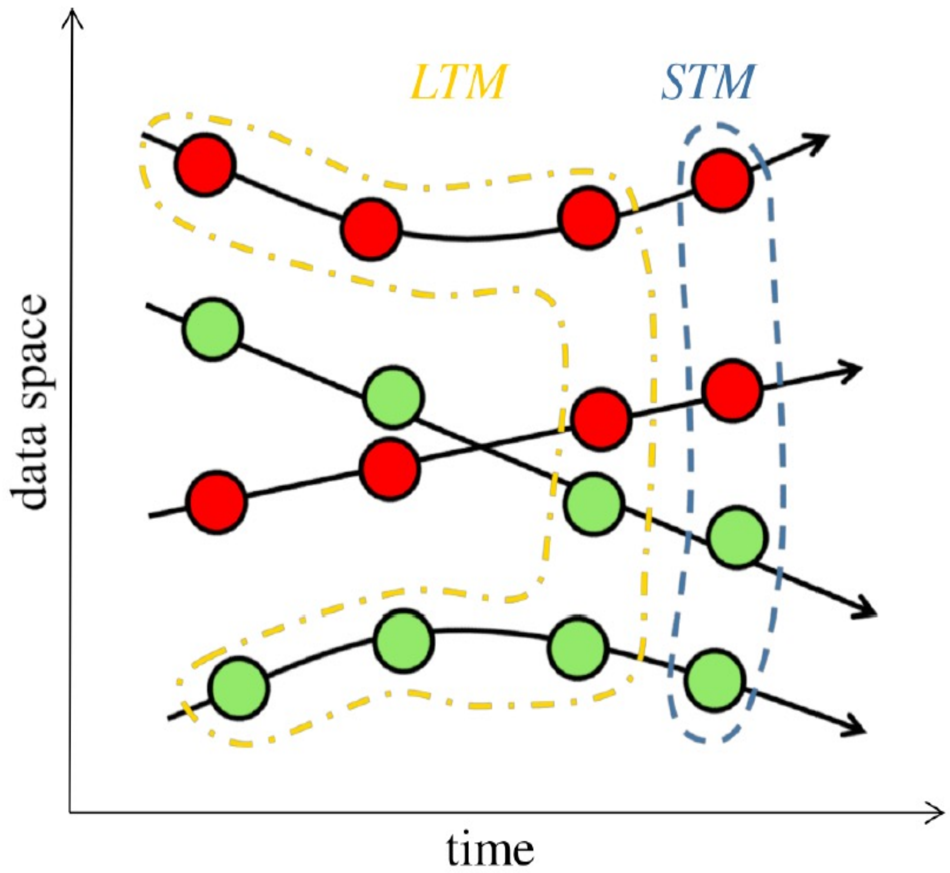
Challenges:

- **Algorithmic challenge** - efficient update for new data point
- **Model selection challenge** - efficient and effective update of model complexity if required
- **Information selection** – forget information which becomes irrelevant due to drift and keep relevant information from the past

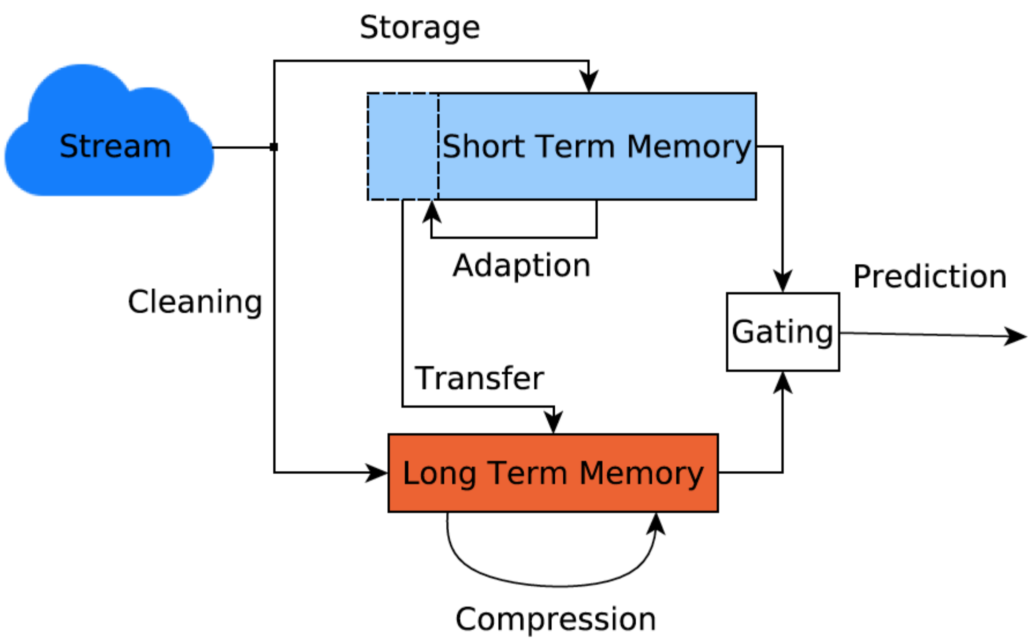
k-NN: basic incremental model

- efficient update
- self-adjusting model-complexity
- unclear: which data to store?

Relevant data



Self-adjusting memory (SAM-kNN)



Parameters:

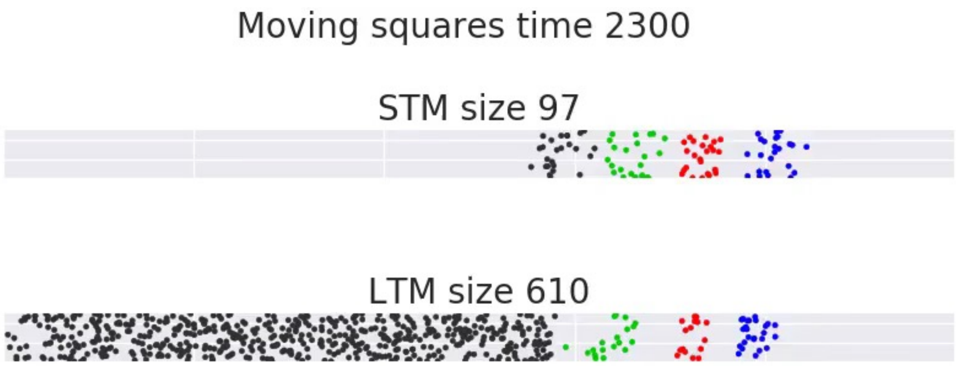
- size of STM
- data points in LTM
- weights of gating

Meta-parameters:

- min size of STM
- max size of STM and LTM
- k of k-NN

Viktor Losing, Barbara Hammer, Heiko Wersing: Tackling heterogeneous concept drift with the Self-Adjusting Memory (SAM). Knowl. Inf. Syst. 54(1): 171-201 (2018), code: https://github.com/vlosing/SAMkNN or within RIVER: https://riverml.xyz/latest/ as SAMKNNClassifier

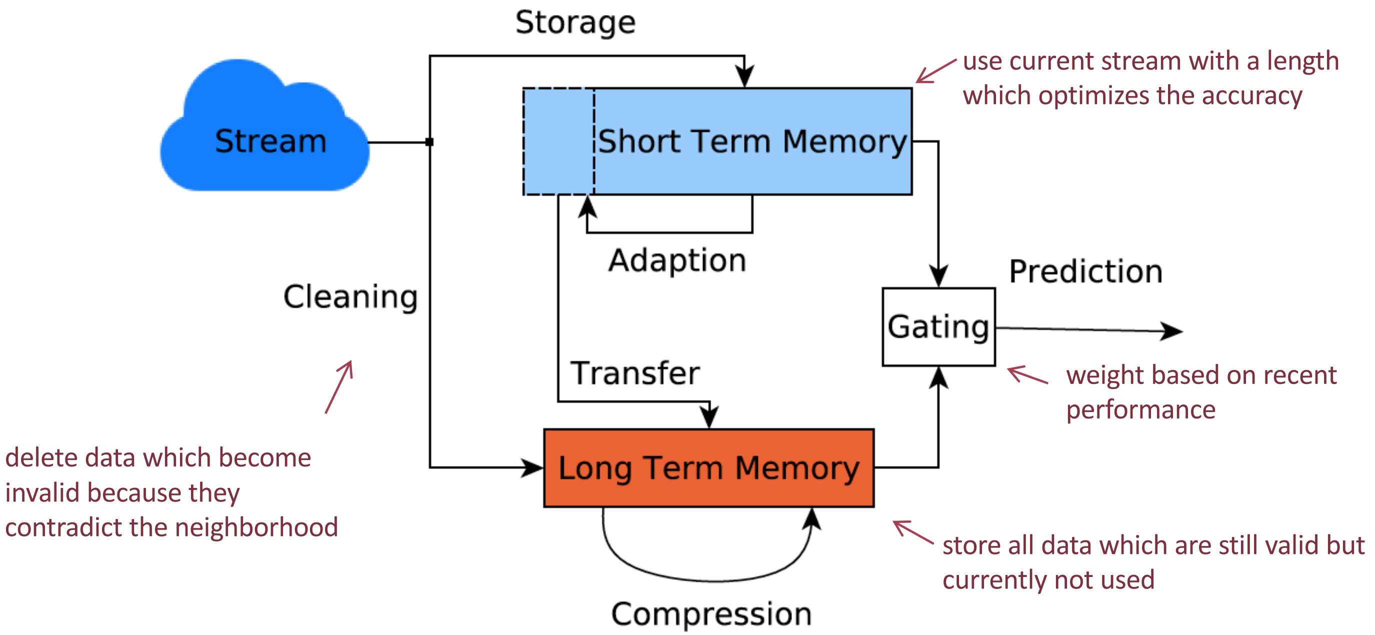
SAM-kNN – example memory



Viktor Losing, Barbara Hammer, Heiko Wersing: Tackling heterogeneous concept drift with the Self-Adjusting Memory (SAM). Knowl. Inf. Syst. 54(1): 171-201 (2018), code: https://github.com/vlosing/SAMkNN or within RIVER: https://riverml.xyz/latest/ as SAMKNNClassifier

UNIVERSITÄT BIELEFELD
Technische Fakultät

Self-adjusting memory (SAM-kNN)



Parameters:

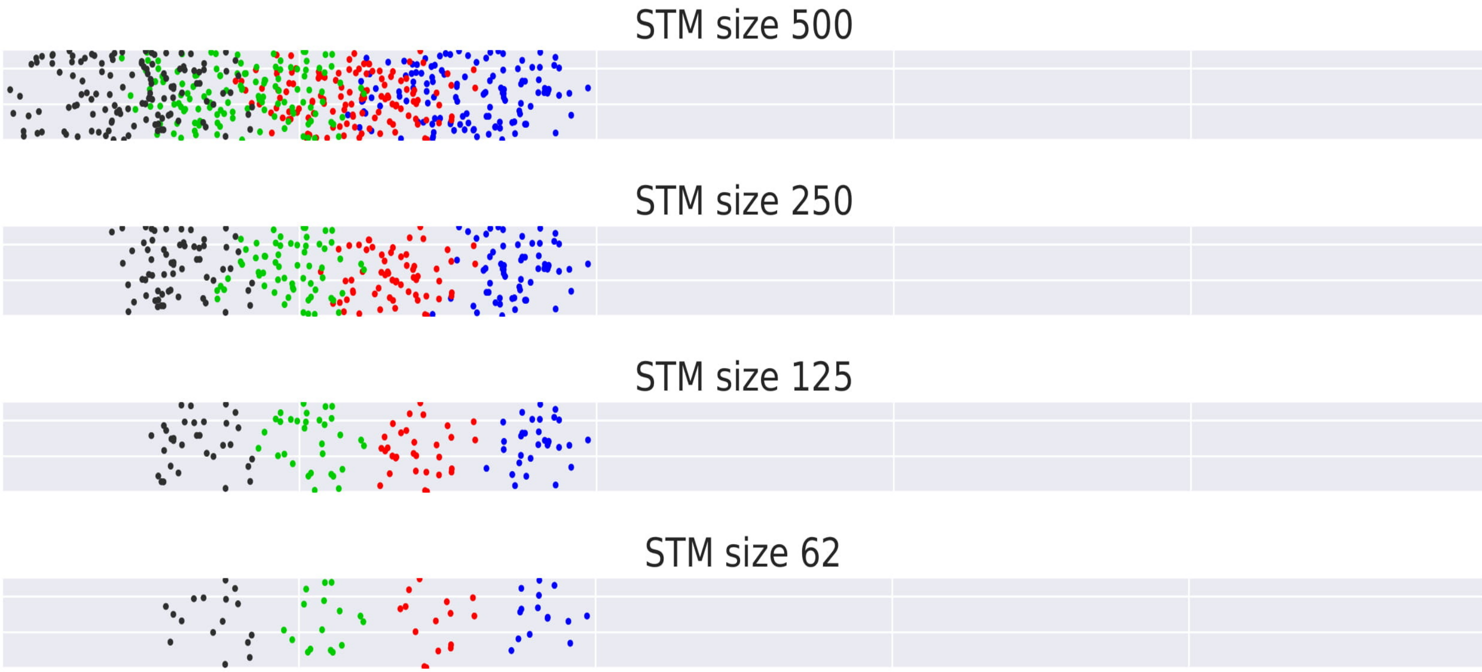
- size of STM
- data points in LTM
- weights of gating

Meta-parameters:

- min size of STM
- max size of STM and LTM
- k of k-NN

Viktor Losing, Barbara Hammer, Heiko Wersing: Tackling heterogeneous concept drift with the Self-Adjusting Memory (SAM). Knowl. Inf. Syst. 54(1): 171-201 (2018), code: https://github.com/vlosing/SAMkNN or within RIVER: https://riverml.xyz/latest/ as SAMKNNClassifier

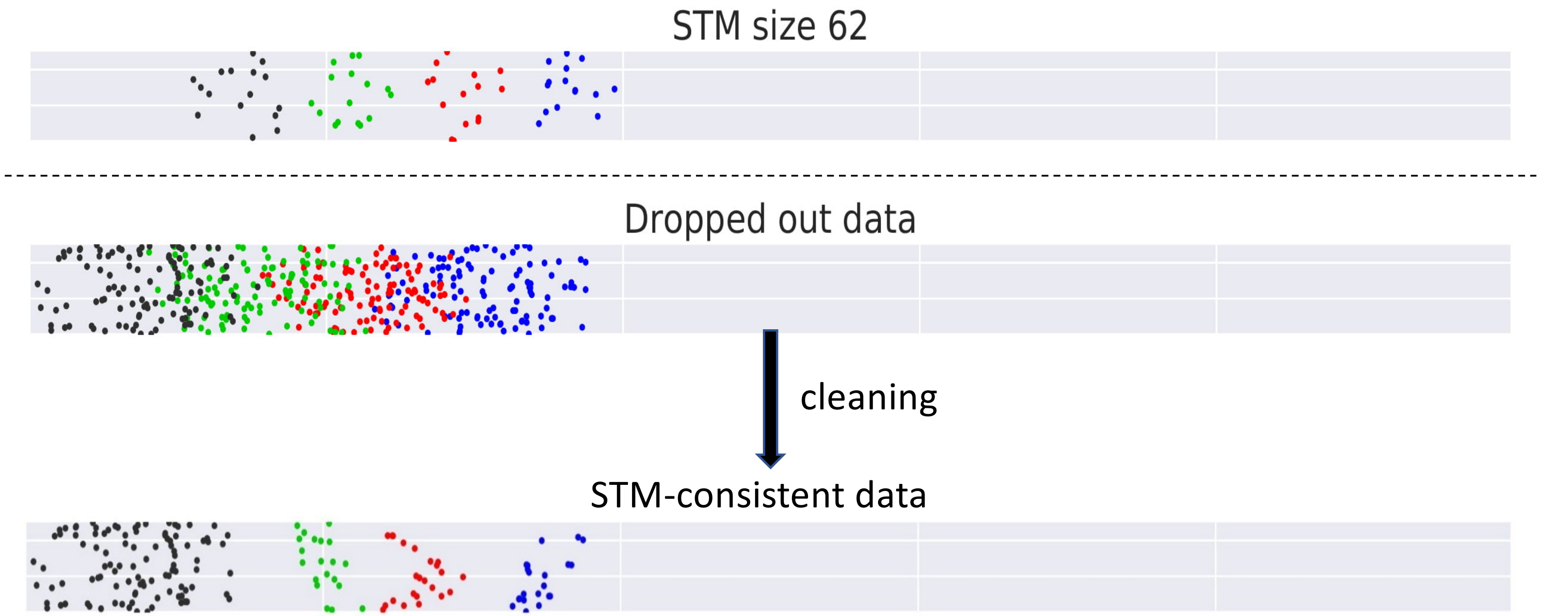
STM adaptation



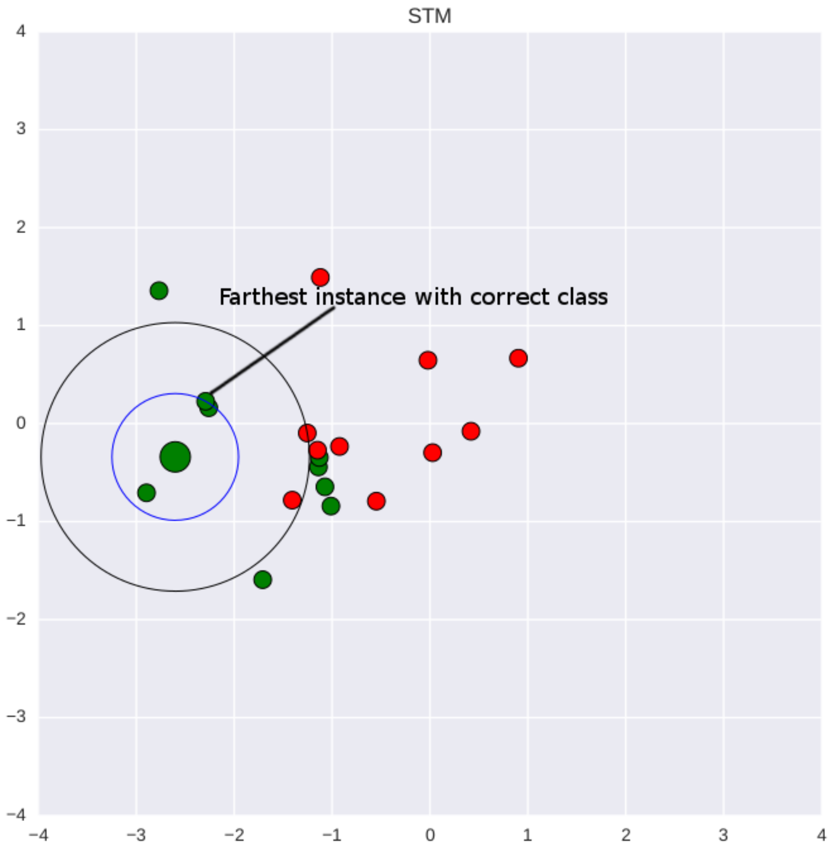
Error
27.12 %
13.12 %
7.12 %
0.0 %

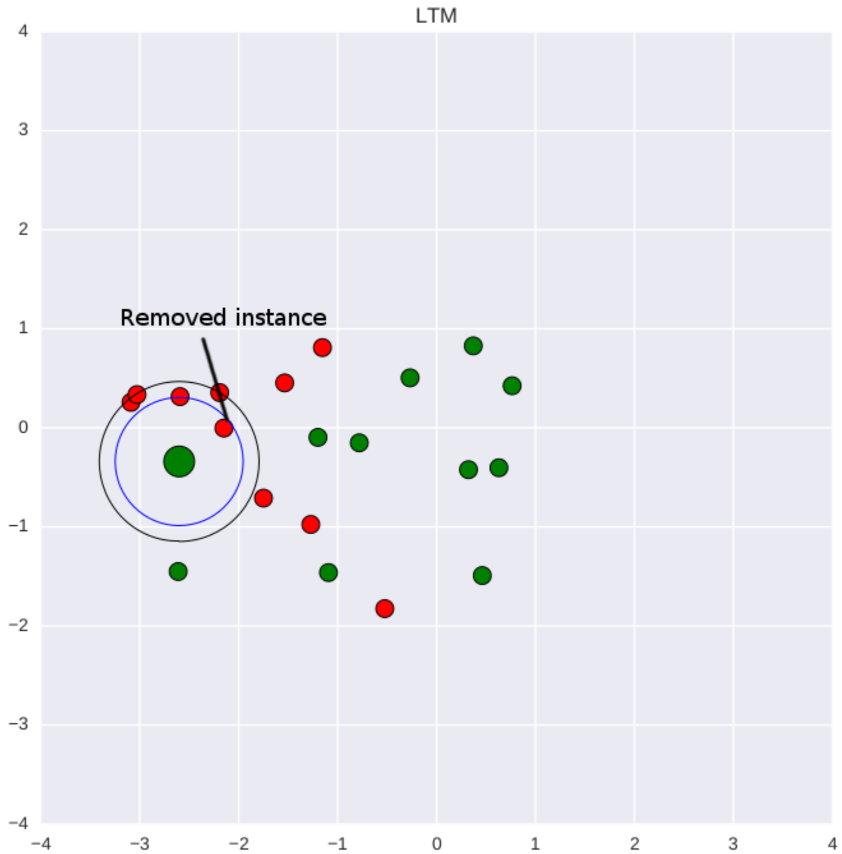
LTM

Transfer consistent data to LTM

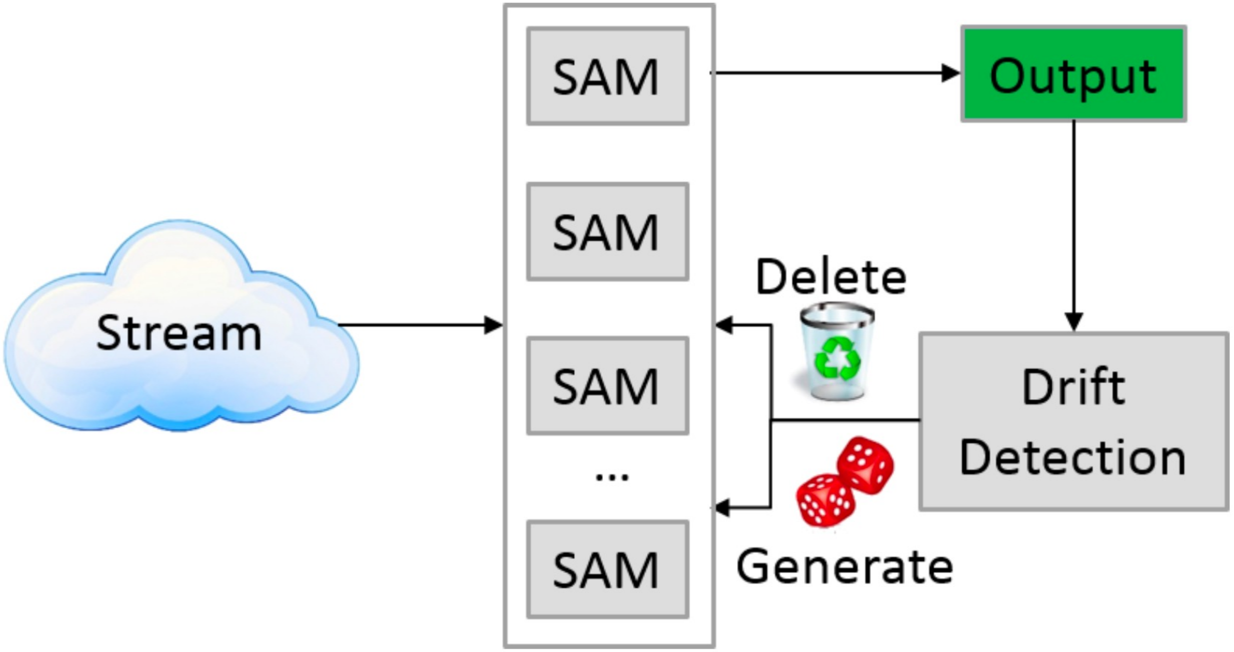


LTM


STM
Farthest instance with correct class
4
3
2
1
0
−1
−2
−3
−4
−4
−3
−2
−1
0
1
2
3
4


LTM
Removed instance
4
3
2
1
0
−1
−2
−3
−4
−4
−3
−2
−1
0
1
2
3
4

Self-adjusting memory ensemble (SAME)



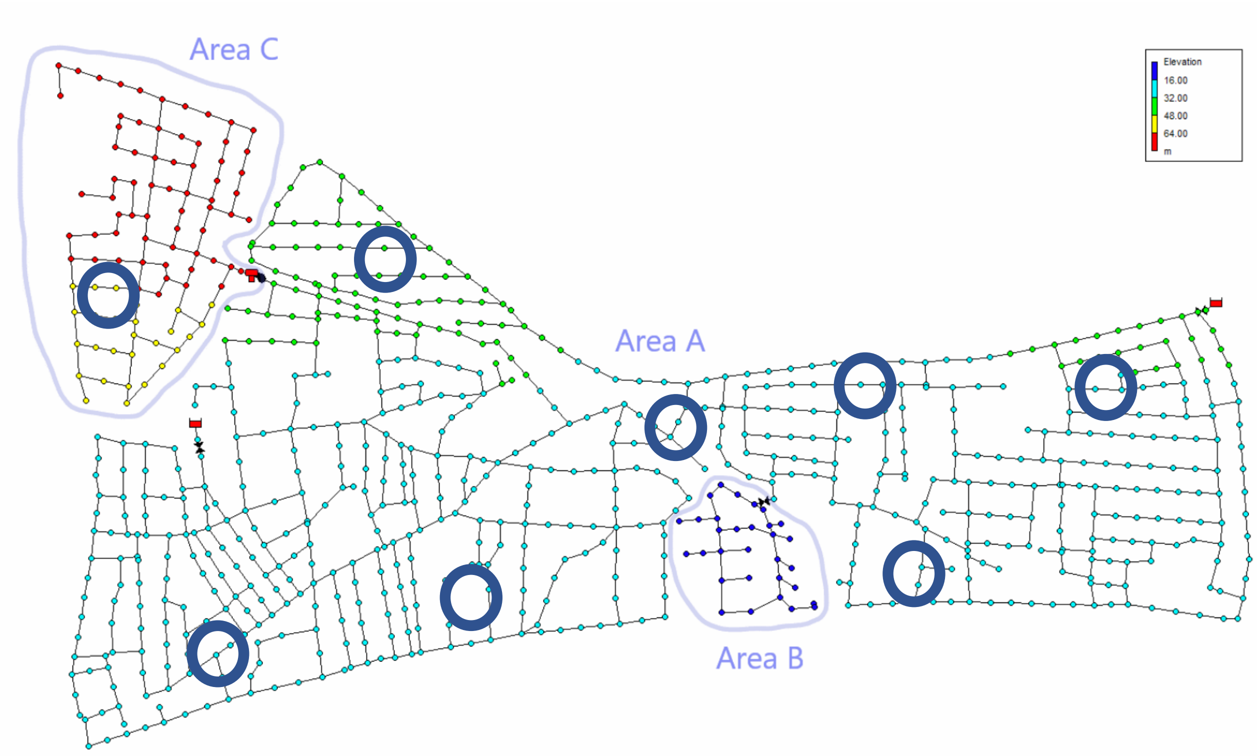
Viktor Losing, Barbara Hammer, Heiko Wersing, Albert Bifet:
Randomizing the Self-Adjusting Memory for Enhanced Handling of Concept Drift. IJCNN 2020: 1-8

Data set	VFDT	SAM	ARF	LVGB	SAM-E
SEA Concepts	15.16	13.22	11.68±0.06	**11.68**±0.07	12.28±0.07
Rot. Hyperplane	15.02	15.22	17.35±0.15	12.73±0.02	**12.49**±0.71
Moving RBF	66.27	12.10	34.02±0.17	45.62±0.15	**11.86**±0.09
Inter. RBF	74.71	3.27	**2.68**±0.04	10.08±0.94	3.30±0.01
Moving Squares	66.73	2.64	36.84±1.49	11.74±0.03	**2.47**±0.25
Transient Chessb.	45.24	11.26	26.30±0.17	14.69±6.22	**10.30**±0.09
Random Tree	10.36	37.05	8.71±1.49	**3.93**±0.09	32.72±0.77
LED-Drift	26.30	45.99	27.39±0.33	**26.13**±0.02	35.48±2.61
Mixed Drift	55.42	12.27	19.87±0.06	25.97±0.10	**11.58**±0.02
Poker	25.88	16.86	19.23±0.17	17.93±0.40	**8.79**±0.44
Artificial ∅	40.11	16.99	20.41	18.05	**14.13**
Outdoor	42.68	11.58	29.70±2.03	39.28±0.25	**9.25**±0.29
Weather	26.49	22.31	21.87±0.46	22.18±0.08	**21.41**±0.16
Electricity	29.00	17.58	21.13±0.50	17.58±0.18	**16.36**±0.19
Rialto	76.19	18.27	24.08±0.10	40.46±0.07	**15.80**±0.16
Airline	34.94	39.84	**34.20**±0.11	36.89±0.02	35.51±0.16
Cover Type	21.85	5.76	8.33±0.03	8.54±0.06	**4.69**±0.36
PAMAP	1.22	0.02	0.03±0.00	0.11±0.01	**0.02**±0.00
SPAM	19.09	7.00	8.18±0.42	7.35±0.31	**5.61**±0.23
KDD99	0.10	0.01	0.03±0.00	0.03±0.00	**0.01**±0.00
Real world ∅	27.95	13.60	16.39	19.16	**12.07**
Overall ∅	34.35	15.38	18.51	18.57	**13.15**
Overall ∅ rank	4.47	2.76	3.00	3.08	**1.68**

Nemenyi significance: SAM-E $\succ$ VFDT

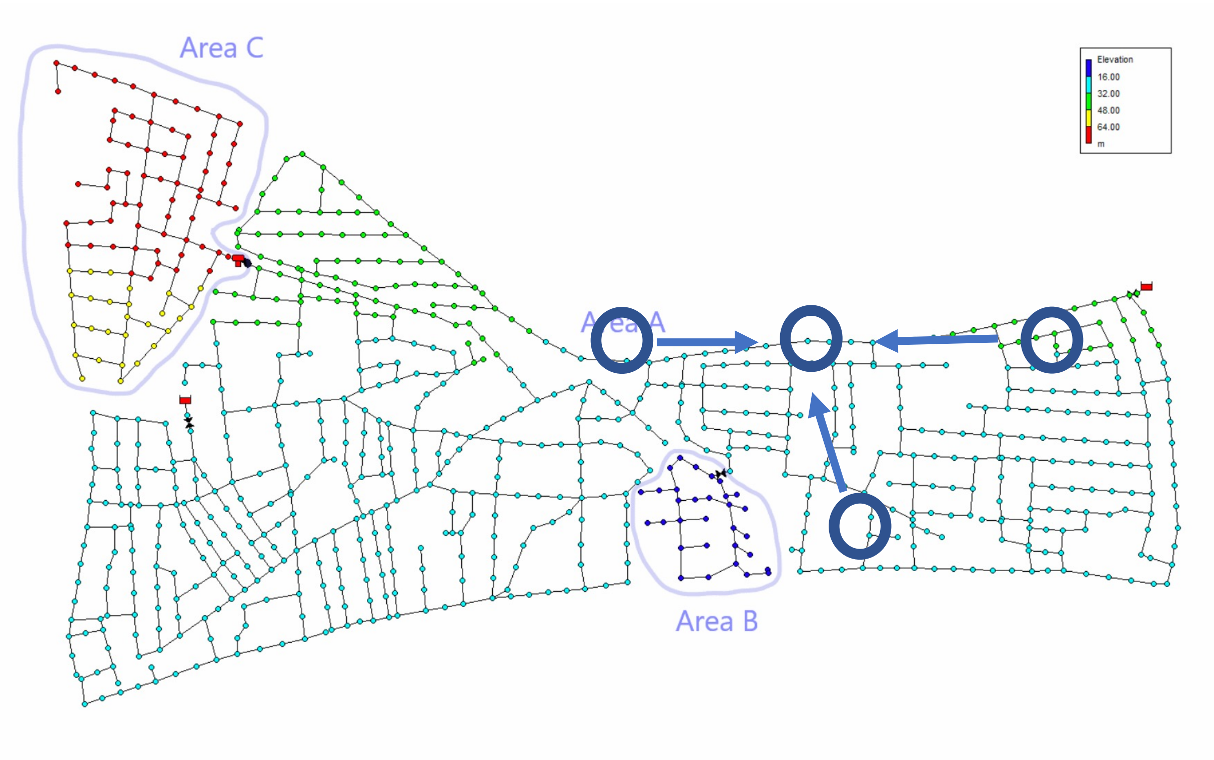
UNIVERSITÄT BIELEFELD
Technische Fakultät

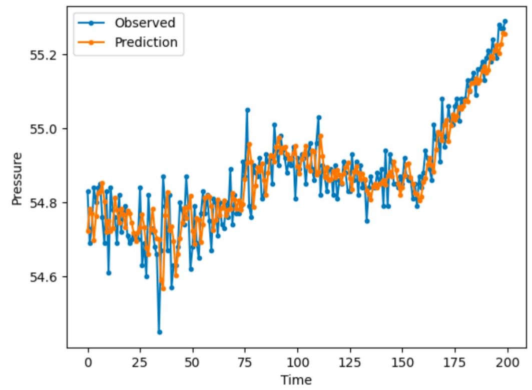
SAM-kNN-regression for fault detection in water distribution systems



UNIVERSITÄT BIELEFELD
Technische Fakultät

Residual-based sensor fault/leakage detection in WDS





pressure sensor

Step 1: predict sensor values from others using incremental time series model
Setp 2: residual based anomaly detection

Performance of SAM-kNN

Method	Metric	Scenario 1	Scenario 2	Scenario 3	Scenario 4	Scenario 5
SAM-kNN	TP	1	1	1	1	1
	FP	48	20	3	20	17
	FN	0	0	0	0	0
kNN	TP	1	1	1	1	1
	FP	17057	19216	11146	19082	18751
	FN	0	0	0	0	0
Linear regression	TP	0	0	0	0	0
	FP	0	0	0	0	0
	FN	1	1	1	1	1

Jonathan Jakob, André Artelt, Martina Hasenjäger, Barbara Hammer:
SAM-kNN Regressor for Online Learning in Water Distribution Networks. ICANN (3) 2022: 752-762

Performance of SAM-kNN

Method	Metric	Scenario 6	Scenario 7	Scenario 8	Scenario 9	Scenario 10
SAM-kNN	TP	1	1	1	1	1
	FP	155	20	157	96	156
	FN	0	0	0	0	0
kNN	TP	1	1	1	1	1
	FP	18596	18596	18596	18596	18596
	FN	0	0	0	0	0
Linear regression	TP	1	0	0	0	0
	FP	0	0	0	0	0
	FN	0	1	1	1	1

Jonathan Jakob, André Artelt, Martina Hasenjäger, Barbara Hammer:
SAM-kNN Regressor for Online Learning in Water Distribution Networks. ICANN (3) 2022: 752-762

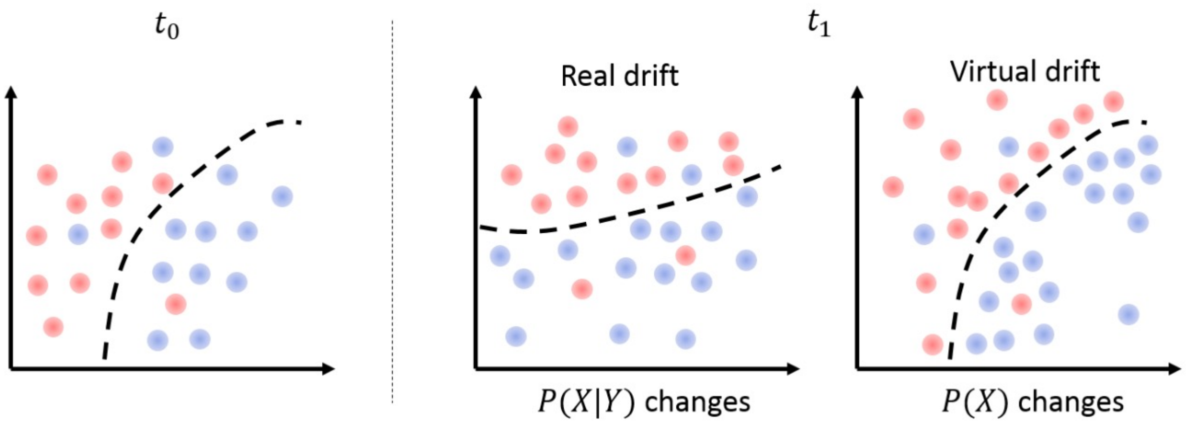
UNIVERSITÄT
BIELEFELD
Technische Fakultät

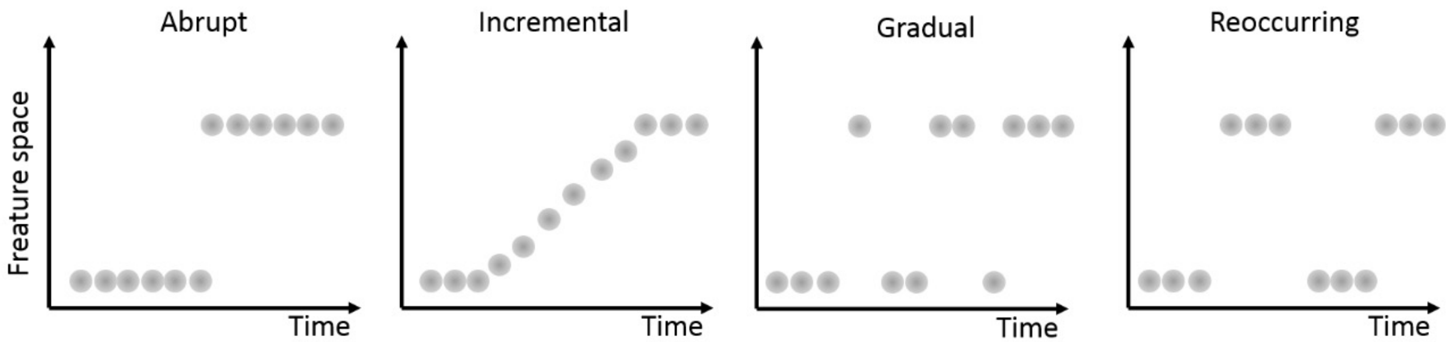
2) Unsupervised scenario: detecting drift

Drift

Drift is present if there exist time points $t_1 \neq t_2$ such that

$$P_{t_1} \neq P_{t_2}$$

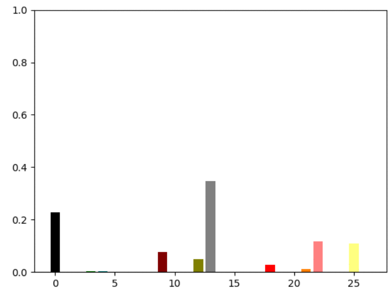




UNIVERSITÄT BIELEFELD

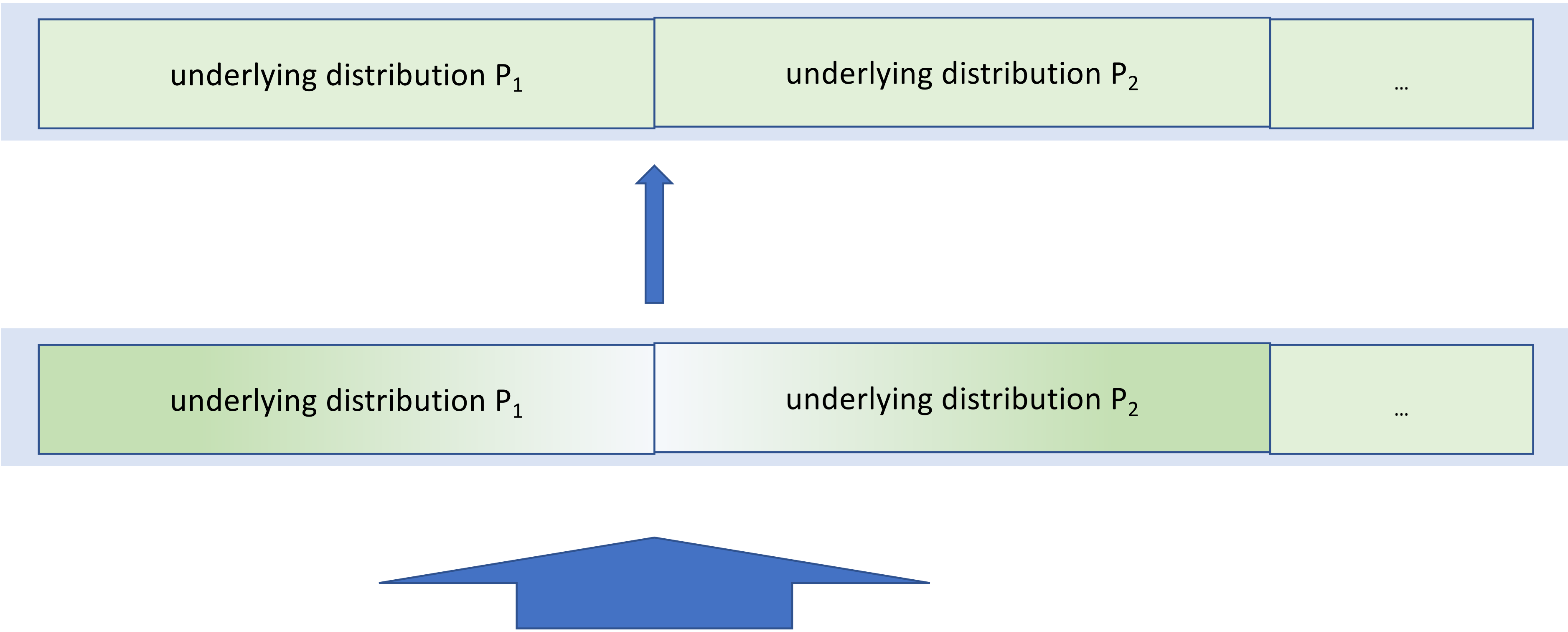
Technische Fakultät

Drift



UNIVERSITÄT BIELEFELD
Technische Fakultät

Drift detection



UNIVERSITÄT BIELEFELD

Technische Fakultät

What is drift?

Drift: data are drawn from a probability distribution P_t which is **not constant** with t

… but we cannot observe P_t

Notions of drift

Drift: data are drawn from a probability distribution P_t which is **not constant** with t

Drift as change of (unobservable) distribution

Dependency of observations and time:

observed data

time

Machine Learner's drift:

optimum model at first time window $\neq$ optimum model at second time window

Towards non-parametric drift detection via Dynamic Adapting Window Independence Drift Detection (DAWIDD), Fabian Hinder, André Artelt, Barbara Hammer, ICML2020

Definition

A *drift process* (p_t, P_T) is a probability measure P_T on $[0,1]$ together with a collection of probability measures p_t on $\mathbb{R}^d$ for all $t \in [0,1]$, such that $t \mapsto p_t(A)$ is measurable for every measurable $A \subset \mathbb{R}^d$, i.e. p_t is a Markov kernel.

Let (p_t, P_T) be a drift process. We say that p_t has *drift* iff $p_t = p_s$ does not hold for P_T-almost all $t, s \in [0,1]$.

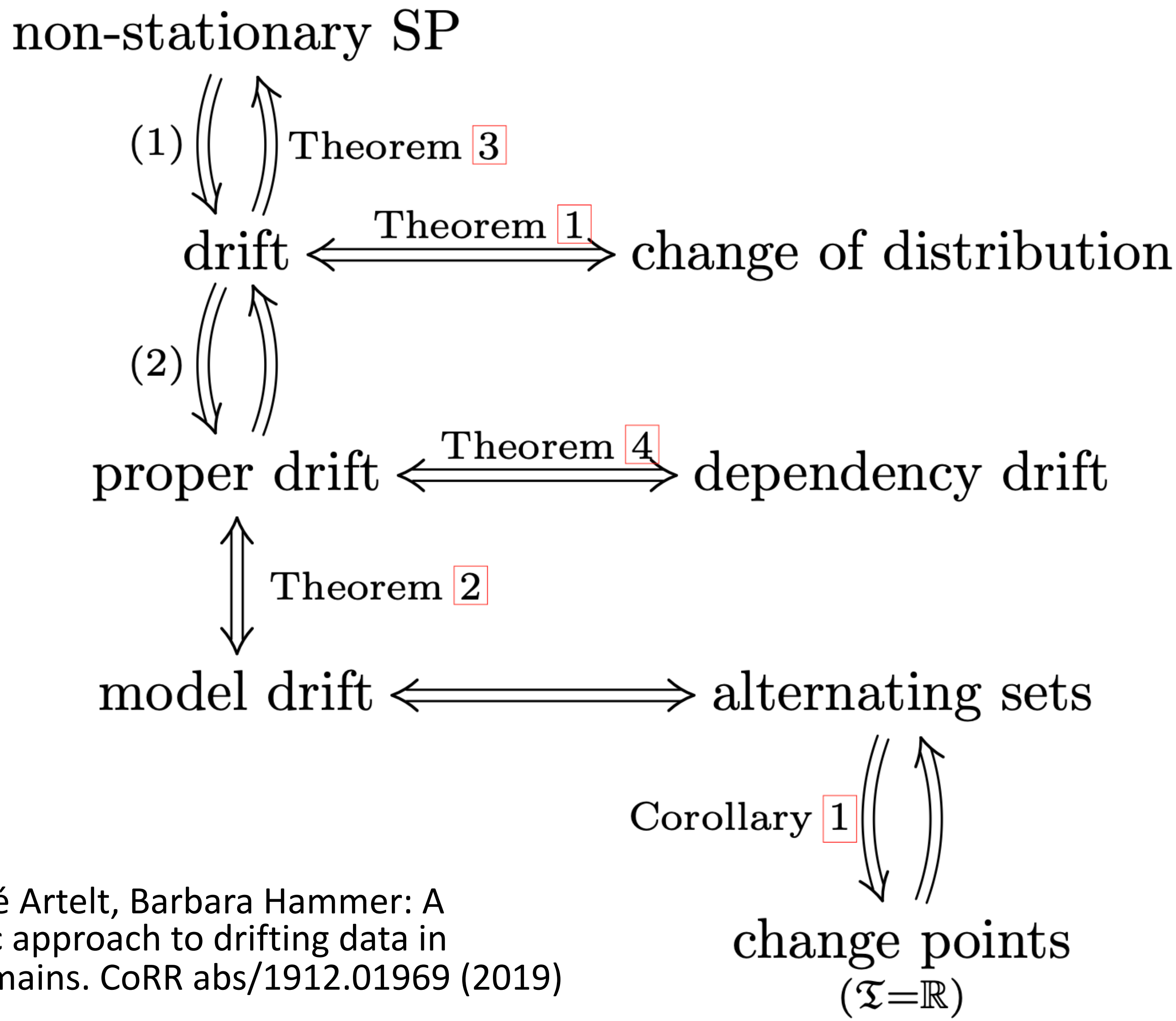
Definition

Let (p_t, P_T) be a drift process and let $(X, T) \sim p_t \otimes P_T$ a pair of random variables. We say that p_t has *dependency drift* iff X and T are statistically dependent, i.e. are not independent random variables.

Definition

We say that a drift process (p_t, P_T) has *model drift* iff there exists measurable sets $A, B \subset [0,1]$ with $P_T(A), P_T(B) > 0$, such that $p_A \neq p_B$, with $p_A = P_T(A)^{-1} \int_A p_t(\cdot) P_T(\mathrm{d}t)$ and analogous for p_B.

UNIVERSITÄT
BIELEFELD
Technische Fakultät

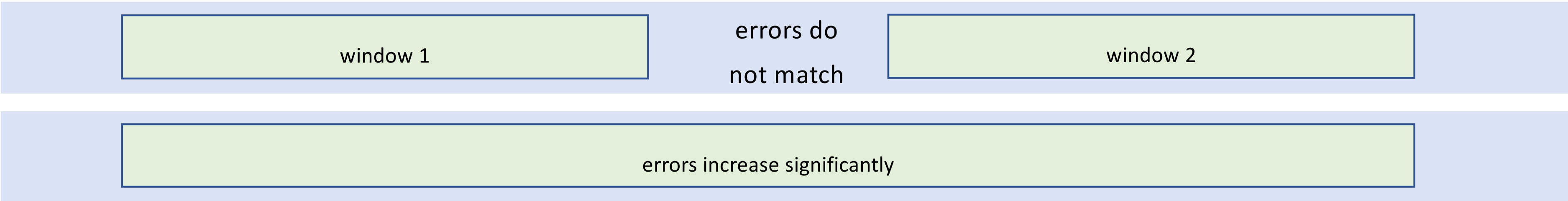


Fabian Hinder, André Artelt, Barbara Hammer: A probability theoretic approach to drifting data in continuous time domains. CoRR abs/1912.01969 (2019)

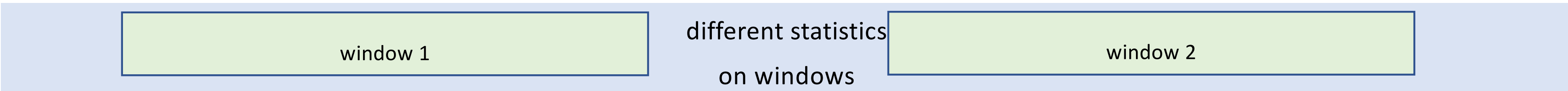
UNIVERSITÄT
BIELEFELD
Technische Fakultät

Drift detection methods

- classification-error based: e.g. ADWIN, DDM, EDDM



- distribution-based: e.g. HDDDM



- dependency based: DAWIDD

UNIVERSITÄT BIELEFELD

Technische Fakultät

DAWIDD

Drift detection as dependency test

Algorithm 1 Dynamic Adaptive Window Independence Drift Detector (DAWIDD)

Input: (x_i) data stream, p p-value for statistical test, $n_{\min}$ minimal number of samples in window, $n_{\max}$ maximal number of samples in window
Initialize Window $W \leftarrow []$
repeat
  Receive new sample x_i at time t_i from stream (x_i)
  $W \leftarrow W \cup \{(x_i, t_i)\}$
  if Test(W, p) rejects H_0 **then**
    output Drift Alert
    Drop $|W| - n_{\min}$ elements from the tail of W
  end if
  while $|W| > n_{\max}$ **do**
    Drop element from W keeping distribution
  end while
until At end of stream (x_i)

Fabian Hinder, André Artelt, Barbara Hammer:
Towards Non-Parametric Drift Detection via Dynamic Adapting Window Independence Drift Detection (DAWIDD). ICML 2020: 4249-4259

DAWIDD

	Dataset	Method	TP	FN	FP	Delay
Real	Weather	DAWIDD	1.4(±0.54)	2.6(±0.54)	6.55(±0.85)	25.0
		HDDDM	0.0	4.0	0.85(±0.13)	–
		EDDM	0.55(±0.25)	3.45(±0.25)	2.55(±0.85)	23.27
		DDM	0.55(±1.15)	3.45(±1.15)	1.7(±2.91)	22.64
		ADWIN	0.15(±0.13)	3.85(±0.13)	1.0(±0.6)	18.0
	Forest Cover Type	DAWIDD	1.4(±0.54)	2.6(±0.54)	7.55(±0.85)	31.82
		HDDDM	0.45(±0.55)	3.55(±0.55)	0.55(±0.25)	28.67
		EDDM	0.4(±0.24)	3.6(±0.24)	2.25(±2.29)	17.38
		DDM	0.3(±0.51)	3.7(±0.51)	1.75(±1.09)	29.5
		ADWIN	0.15(±0.13)	3.85(±0.13)	2.3(±1.71)	29.0
	Electricity Market	DAWIDD	0.15(±0.13)	3.85(±0.13)	1.3(±2.01)	21.0
		HDDDM	0.0	4.0	0.1(±0.09)	–
		EDDM	0.3(±0.21)	3.7(±0.21)	2.5(±0.75)	31.0
		DDM	1.2(±1.56)	2.8(±1.56)	2.85(±1.93)	20.42
		ADWIN	0.4(±0.34)	3.6(±0.34)	2.4(±1.14)	23.88

Fabian Hinder, André Artelt, Barbara Hammer: Towards Non-Parametric Drift Detection via Dynamic Adapting Window Independence Drift Detection (DAWIDD). ICML 2020: 4249-4259

UNIVERSITÄT
BIELEFELD
Technische Fakultät

DAWIDD

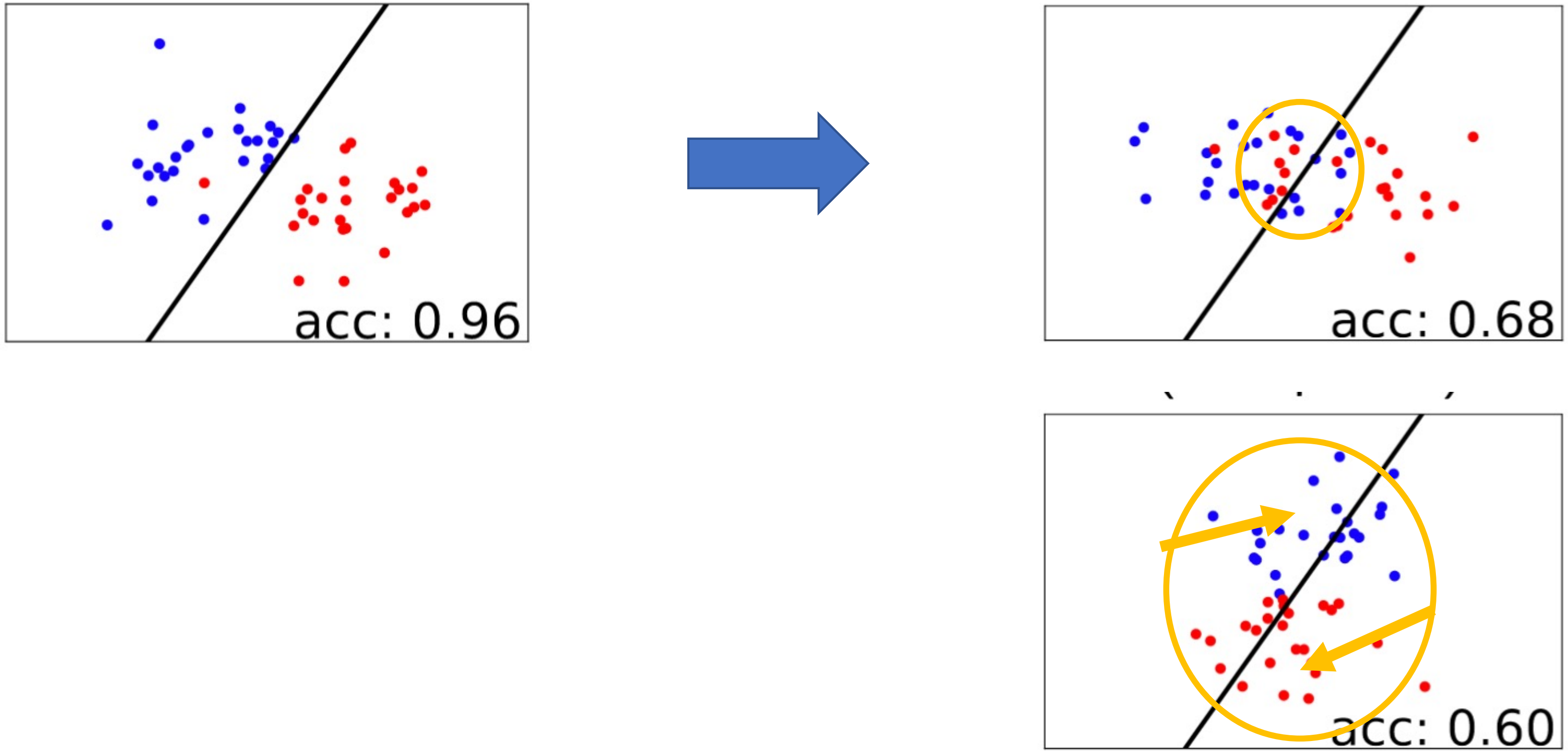
Method	TP	FN	FP	Delay
DAWIDD	**2.2**	**2.2**	3.3	**2.4**
HDDDM	3.4	3.4	**1.8**	3.3
EDDM	2.9	2.9	3.3	2.9
DDM	2.9	2.9	3.3	2.8
ADWIN	3.6	3.6	3.3	3.5

Fabian Hinder, André Artelt, Barbara Hammer:
Towards Non-Parametric Drift Detection via Dynamic Adapting Window Independence Drift Detection (DAWIDD). ICML 2020: 4249-4259

UNIVERSITÄT
BIELEFELD
Technische Fakultät

3) Unsupervised scenario: Explaining drift

Drift explanation



UNIVERSITÄT
BIELEFELD

Technische Fakultät

Towards drift explanation

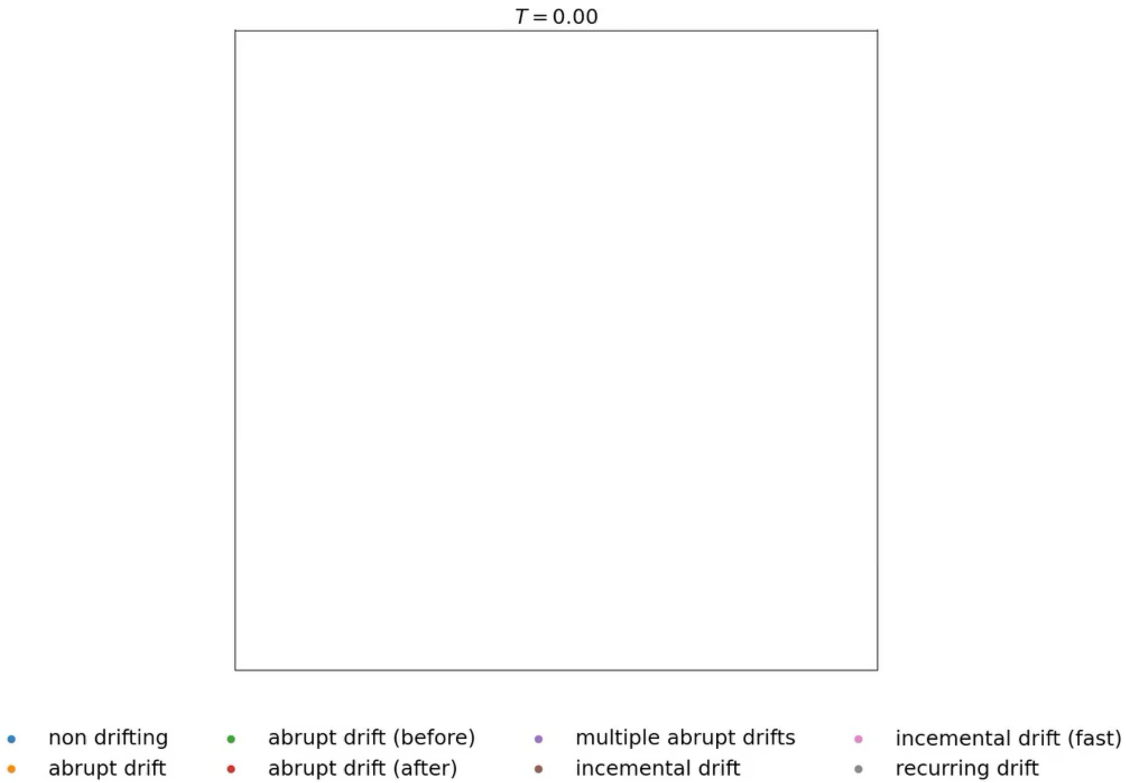
Assume drift is present, i.e. $t_1 \neq t_2$ such that $P_{t_1} \neq P_{\mathrm{t}_2}$

Drift localization:

Identify regions D of the data space such that

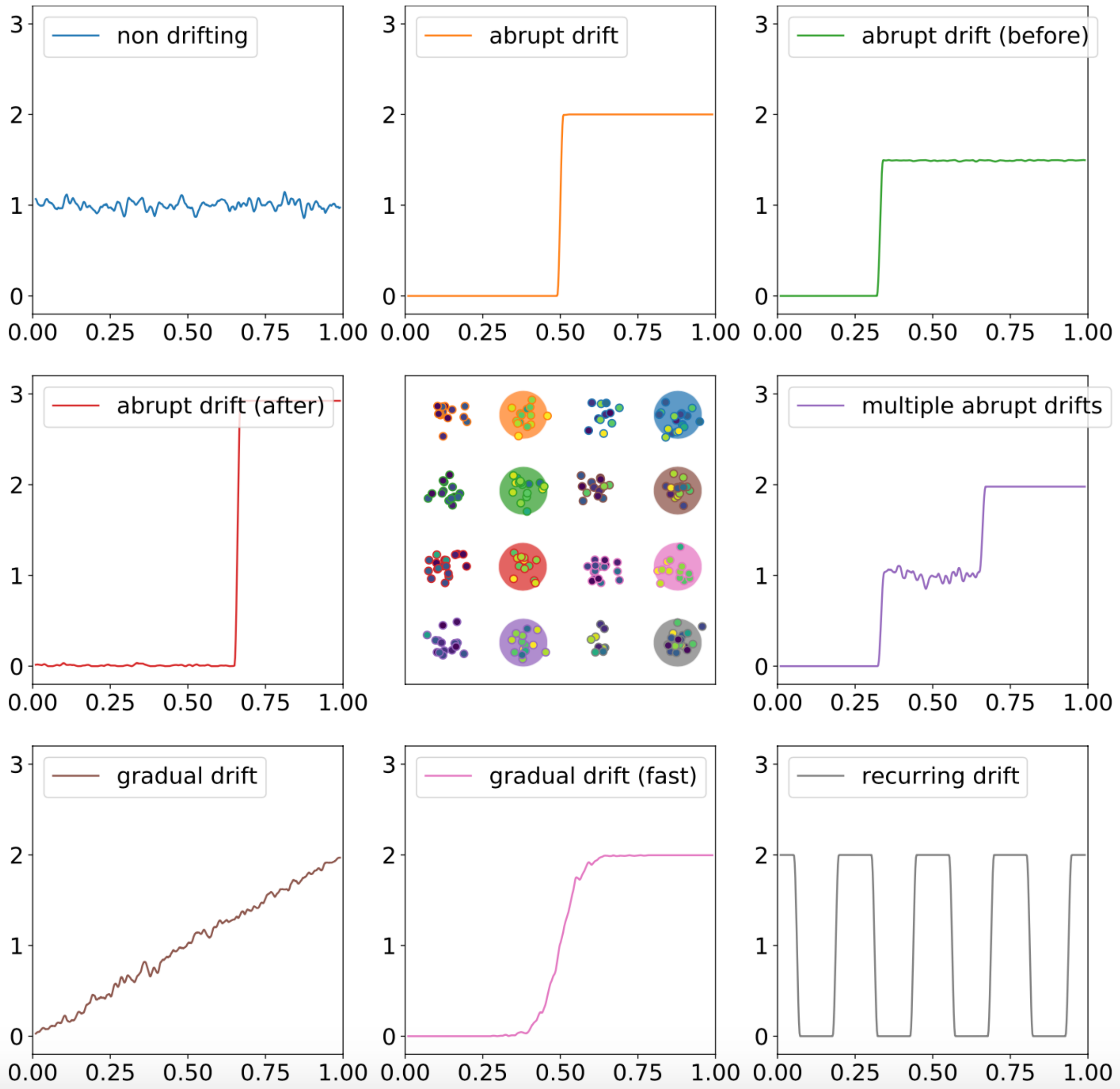
$P_{t_1}(D) \neq P_{\mathrm{t}_2}(D)$ and $P_{t_1}(D^c) = P_{\mathrm{t}_2}(D^c)$

Where is drift?



UNIVERSITÄT BIELEFELD

Technische Fakultät



UNIVERSITÄT BIELEFELD
Technische Fakultät

Drift segmentation

Assume drift is present, i.e. $t_1 \neq t_2$ such that $P_{t_1} \neq P_{\mathrm{t}_2}$

Drift localization:

Identify regions D of the data space such that

$P_{t_1}(D) \neq P_{\mathrm{t}_2}(D)$ and $P_{t_1}(D^c) = P_{\mathrm{t}_2}(D^c)$

Drift segmentation:

Find a segmentation function $L: X \to \mathbb{N}$ with small $|L(X)|$

such that $L(x) = L(x') \Rightarrow P(T|X = x) = P(T|X = x')$

Drift segmentation

Lemma 1. *Let* $L : \mathcal{X} \to \mathbb{N}$. *Then* L *is a drift segmentation if and only if* T *and* X *are independent given* $L(X)$, *i.e.* $T \perp\!\!\!\perp X | L(X)$.

Algorithm: decision tree algorithm where splits into subsets l_1 and l_2 are chosen such that the difference of $P(T|l_1)$ and $P(T|l_2)$ is maximum

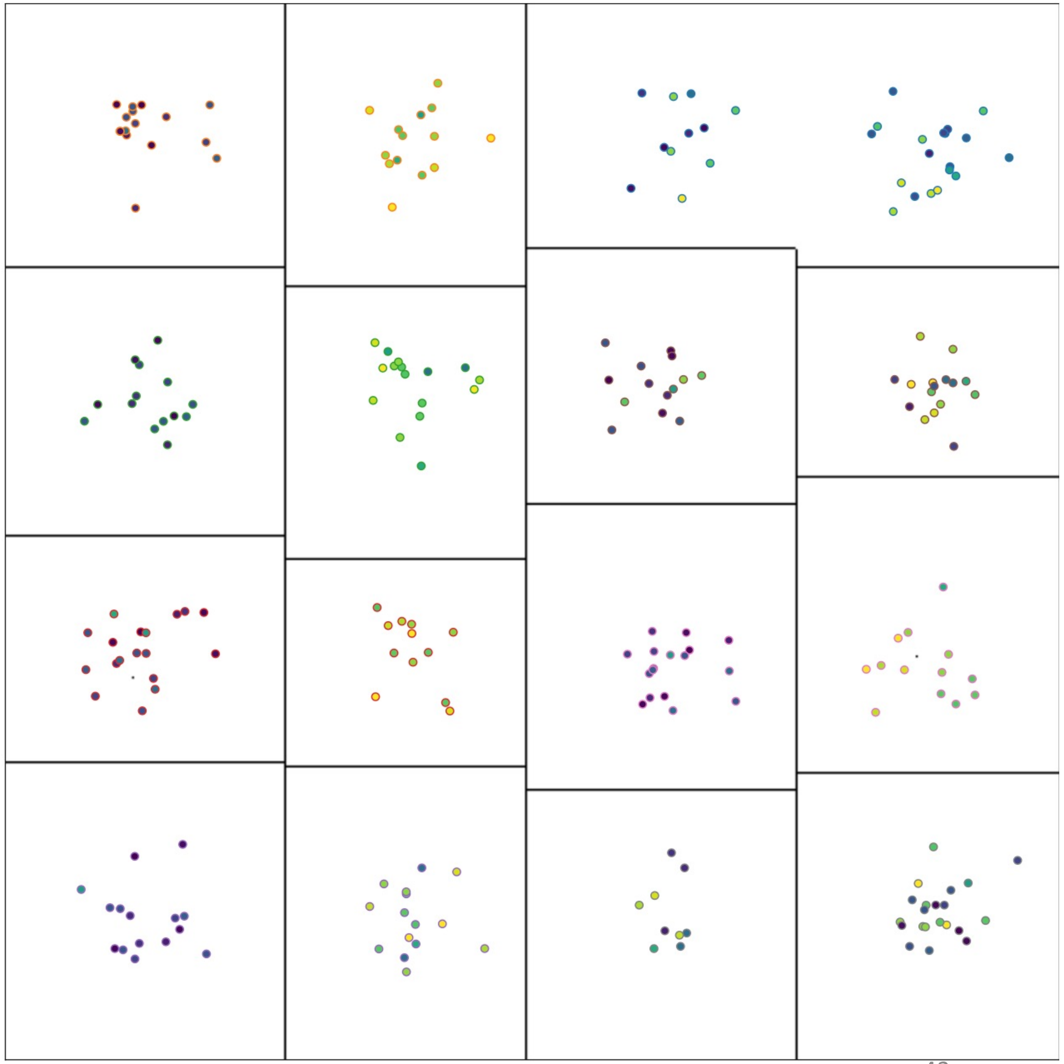
Use Kolmogorov-Smirnov Test statistics for sets of points ordered according to time:

$$\left\| \hat{F}_{T|X\in l_1} - \hat{F}_{T|X\in l_2} \right\|_\infty = \max_{1\leq k\leq N} \left| \frac{k}{n} - \frac{N}{n\cdot(N-n)} \sum_{i=1}^{k} \mathbb{I}_{l_2}(x_i) \right|$$

UNIVERSITÄT BIELEFELD

Technische Fakultät

Drift segmentation



UNIVERSITÄT BIELEFELD
Technische Fakultät

Evaluation w.r.t drift localization

Table: Experimental results over 200 runs. Mean accuracy and standard deviation are shown. Significantly ($p = 0.01$) better results are printed in bold face. n is the number of noise dimensions, cpt is the number of clusters per time.

cpt	n	Kolmogorov	k-NN	LDD-DSI	kdq-Tree
9	0	**0.87(±0.09)**	**0.86(±0.07)**	0.60(±0.03)	0.78(±0.11)
9	1	**0.86(±0.11)**	0.75(±0.07)	0.49(±0.06)	0.70(±0.09)
18	0	0.73(±0.09)	**0.78(±0.05)**	0.60(±0.03)	0.72(±0.08)
18	1	**0.74(±0.09)**	0.69(±0.04)	0.48(±0.06)	0.66(±0.06)
18	5	**0.71(±0.10)**	0.58(±0.01)	0.37(±0.02)	0.48(±0.05)

F.Hinder, B.Hammer, Concept drift segmentation via Kolmogorov trees, ESANN21

UNIVERSITÄT BIELEFELD
Technische Fakultät

Evaluation w.r.t conditional density estimation

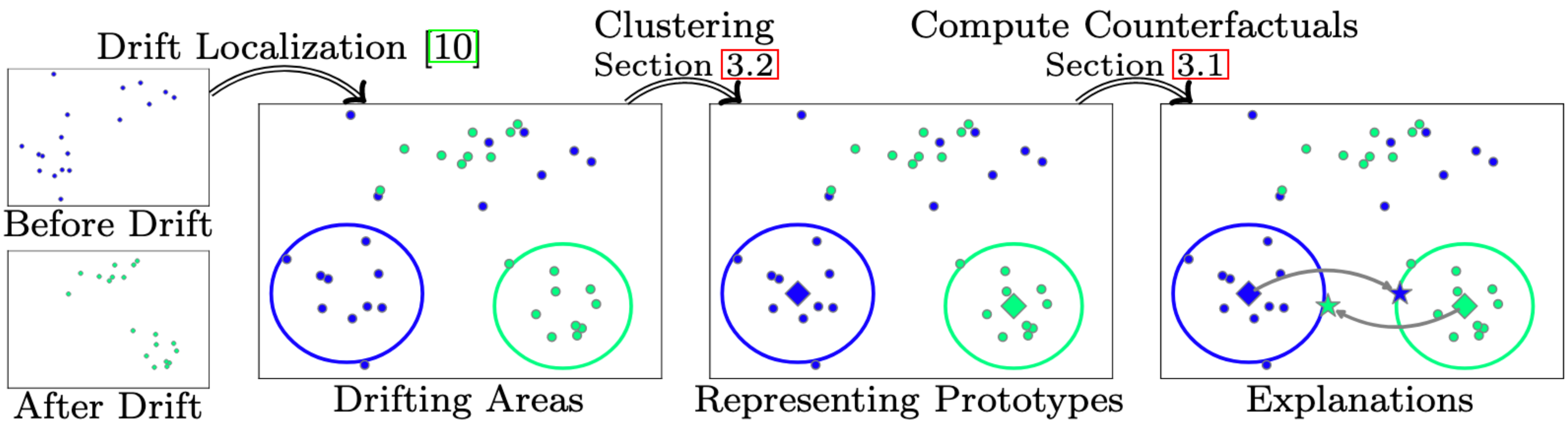
Table: Experimental results over 200 runs. Table shows mean negative log-likelihood and standard deviation. Significantly ($p = 0.01$) better results are printed in bold face. Number in brackets denotes the number of "pearls".

	Kolmogorov	LS-CDE	MSE	ϵ-KDE
boston	$\mathbf{0.45}(\pm\mathbf{0.04})$	$0.65(\pm 0.10)$	$\mathbf{0.44}(\pm\mathbf{0.06})$	$1.17(\pm 0.05)$
california housing	$0.83(\pm 0.03)$	$0.89(\pm 0.04)$	$\mathbf{0.74}(\pm\mathbf{0.04})$	$1.05(\pm 0.03)$
diabetes	$1.11(\pm 0.03)$	$1.18(\pm 0.05)$	$\mathbf{1.08}(\pm\mathbf{0.04})$	$1.73(\pm 0.05)$
Gauss necklace (3)	$\mathbf{1.25}(\pm\mathbf{0.03})$	$1.29(\pm 0.04)$	$1.31(\pm 0.04)$	$1.46(\pm 0.05)$
Gauss necklace (6)	$\mathbf{1.22}(\pm\mathbf{0.02})$	$1.25(\pm 0.03)$	$1.31(\pm 0.04)$	$1.43(\pm 0.04)$

F.Hinder, B.Hammer, Concept drift segmentation via Kolmogorov trees, ESANN21

UNIVERSITÄT BIELEFELD
Technische Fakultät

Drift explanation



Fabian Hinder, André Artelt, Valerie Vaquet, Barbara Hammer:
Contrasting explanations of concept drift, ESANN 2022

Drift explanation

Drift explanation algorithm:

- detect drift (‘when’ - e.g. using DAWIDD)
- detect region in space D where drift is present (‘where’ - e.g. using Kolmogorov trees)
- learn a model h which maps the regions of drift to the time ‘before’ / ‘after’ / ‘either’ (e.g. standard decision trees)
- collect representatives x of drifting regions (e.g. using affinity propagation)
- use contrasting explanation to explain h w.r.t. x in D (e.g. using CEML toolbox: https://github.com/andreArtelt/ceml)

Fabian Hinder, André Artelt, Valerie Vaquet, Barbara Hammer: Contrasting explanations of concept drift, ESANN 2022

UNIVERSITÄT
BIELEFELD
Technische Fakultät

Evaluation of drift explanation

Scenario I:

induce feature drift for
Nebraska weather data

Explain by
sparse counterfactuals

FP	LE	precision	recall	F1
gaussian	0%	0.84±0.31	0.91±0.29	0.87±0.29
	10%	0.84±0.33	0.89±0.31	0.85±0.32
	20%	0.80±0.35	0.88±0.32	0.82±0.33
	40%	0.78±0.37	0.85±0.36	0.80±0.36
shift	0%	0.86±0.24	0.99±0.10	0.90±0.17
	10%	0.88±0.22	1.00±0.00	0.92±0.14
	20%	0.86±0.25	0.98±0.14	0.90±0.19
	40%	0.87±0.25	0.97±0.17	0.90±0.21
zero	0%	1.00±0.00	1.00±0.00	1.00±0.00
	10%	1.00±0.00	1.00±0.00	1.00±0.00
	20%	1.00±0.00	1.00±0.00	1.00±0.00
	40%	1.00±0.00	1.00±0.00	1.00±0.00

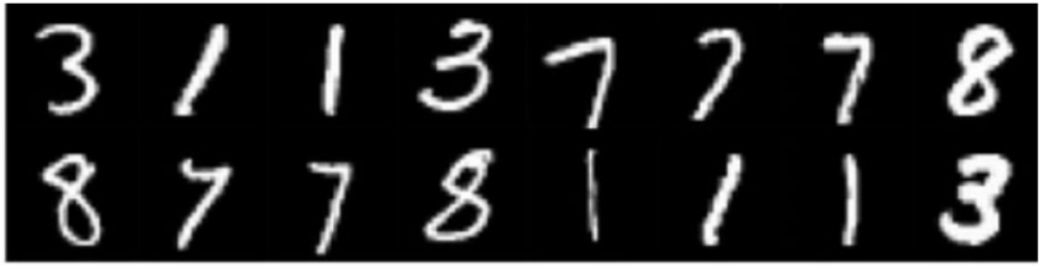
Evaluation of drift explanation

Scebario II:

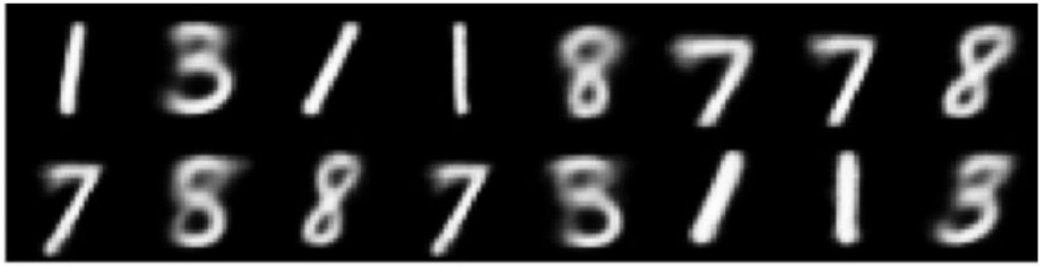
MNIST data with

classes 1,3,4 (before)

and 7,8,4 (after)



(a) Explanation using raw data.



(b) Explanation using VAE.

UNIVERSITÄT
BIELEFELD
Technische Fakultät

Example of drift explanation

UNIVERSITÄT
BIELEFELD
Technische Fakultät

Conclusions

Conclusions

- supervised learning on streaming data, e.g. using SAM-kNN or other methods from River toolbox: https://riverml.xyz/0.14.0/
- drift characterization as dependency X and T
- drift detection based on dependence test: https://github.com/FabianHinder/DAWIDD
- drift segmentation / localization based on difference of $P(T|L(x))$
- drift explanation based on contrasting explanations: https://github.com/FabianHinder/Contrasting-Explanation-of-Concept-Drift

UNIVERSITÄT BIELEFELD

Technische Fakultät

Thanks

joint work with:

André Artelt (Uni Bielefeld), Albert Bifet (Uni Télécom Paris), Martina Hasenjäger (HRI Europe), Fabian Hinder (Uni Bielefeld), Jonathan Jakob (Uni Bielefeld), Viktor Losing (CrowdStrike), Heiko Wersing (HRI Europe), Valerie Vaquet (Uni Bielefeld), Taizo Yoshikawa (Honda R&D Co)

erc

VolkswagenStiftung

Bundesministerium für Bildung und Forschung

DFG

HRI Honda Research Institute EU

Bundesministerium für Wirtschaft und Energie

Ministerium für Kultur und Wissenschaft des Landes Nordrhein-Westfalen