Cognitive Interaction Technology
Center of Excellence
Bielefeld University

Honda Research Institute Europe

Memory Models for Incremental Learning Architectures

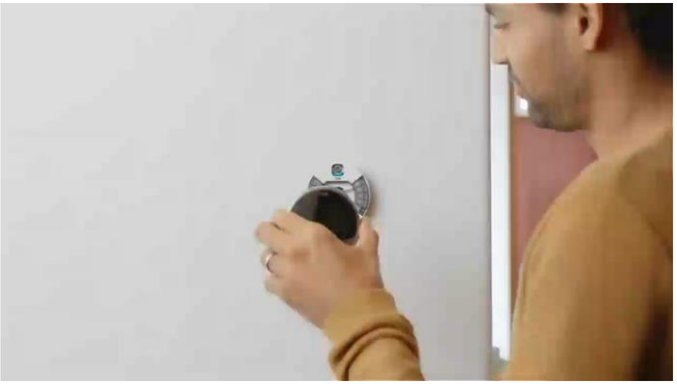
Viktor Losing, Heiko Wersing and Barbara Hammer

Outline

- Motivation
- ~~Case study: Personalized Maneuver Prediction at Intersections~~
- Handling of Heterogeneous Concept Drift

Motivation

- Personalization
  – adaptation to user habits / environments
- Lifelong-learning

Challenges - Personalized online learning

- Learning from few data

Challenges - Personalized online learning

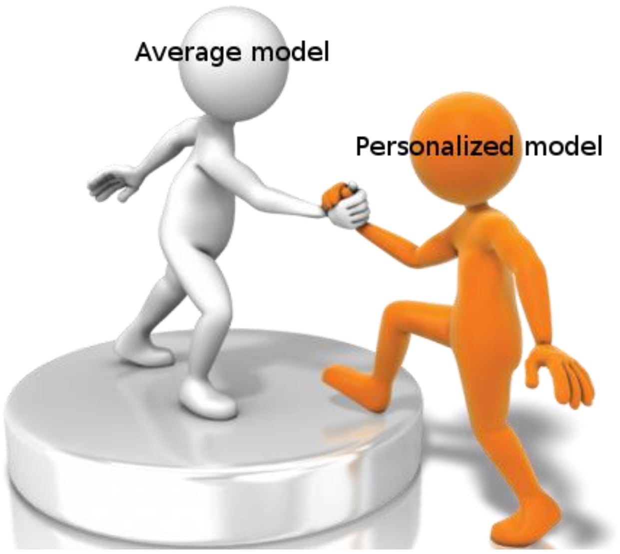
- Learning from few data
- Sequential data with predefined order

Challenges - Personalized online learning

- Learning from few data
- Sequential data with predefined order
- Concept drift

Challenges - Personalized online learning

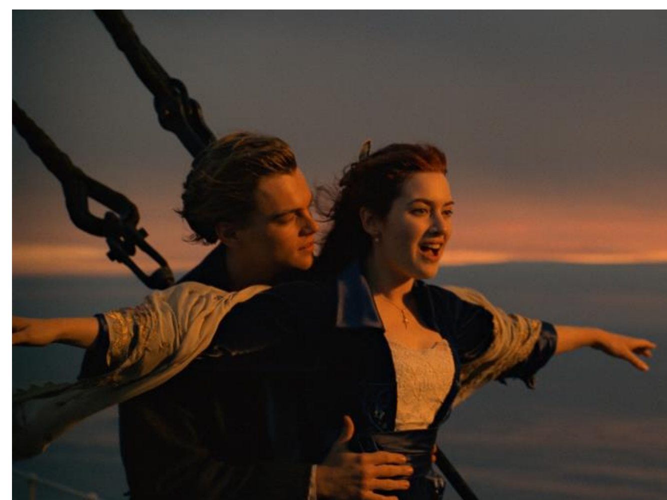
- Learning from few data
- Sequential data with predefined order
- Concept drift
- Cooperation between average and personalized model



Change is everywhere

- Coping with „arbitrary“ changes

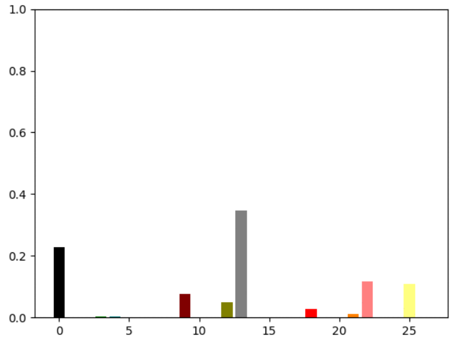
Change of taste / interest



Seasonal changes

Change of context

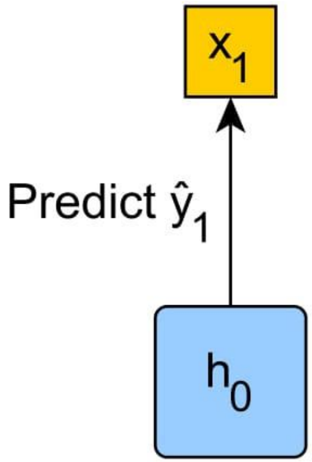
Rialto task: Change of lighting conditions



Setting

- Supervised stream classification
  - Predict for an incoming stream of features $x_1, \ldots, x_j, x_i \; \mathbb{R}^n$ the corresponding labels $y_1, \ldots y_j, y_i \in \{1, \ldots, c\}$

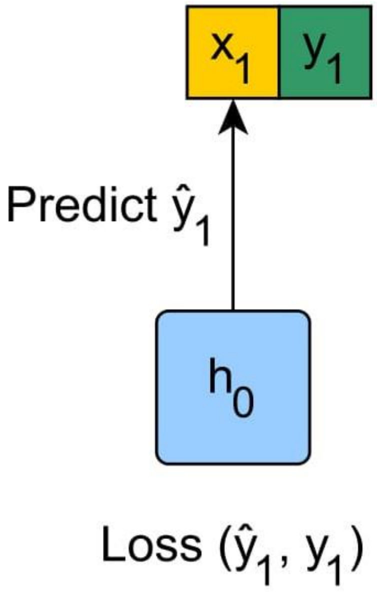
- On-line learning scheme
  - After each touple (x_i, y_i)generate a new model h_i to predict the next incoming example



Setting

- Supervised stream classification
  - Predict for an incoming stream of features $x_1, \ldots, x_j, x_i\ \mathbb{R}^n$ the corresponding labels $y_1, \ldots y_j, y_i \in \{1, \ldots, c\}$

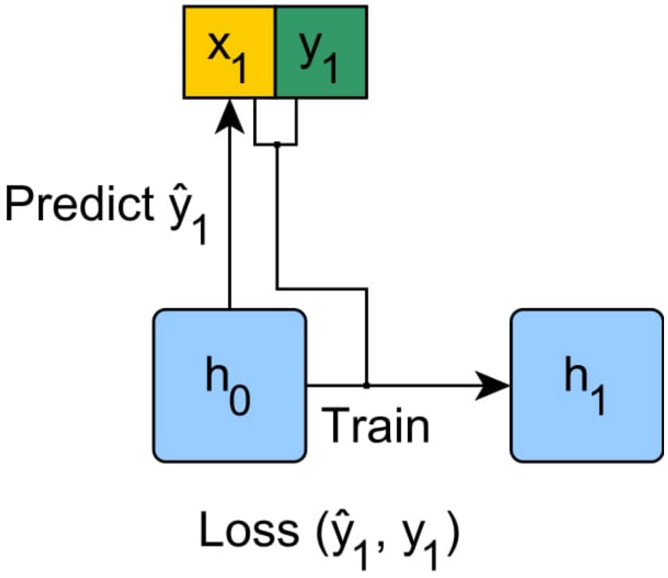
- On-line learning scheme
  - After each touple (x_i, y_i)generate a new model h_i to predict the next incoming example



Setting

- Supervised stream classification
  - Predict for an incoming stream of features $x_1, \dots, x_j, x_i\ \mathbb{R}^n$ the corresponding labels $y_1, \dots y_j, y_i \in \{1, \dots, c\}$

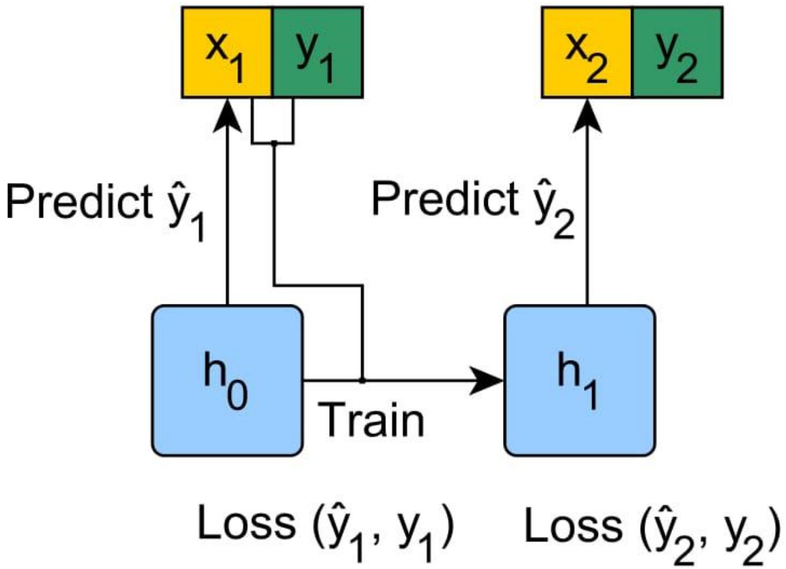
- On-line learning scheme
  - After each touple (x_i, y_i)generate a new model h_i to predict the next incoming example



Setting

- Supervised stream classification
  - Predict for an incoming stream of features $x_1, \ldots, x_j, x_i \ \mathbb{R}^n$ the corresponding labels $y_1, \ldots y_j, y_i \in \{1, \ldots, c\}$

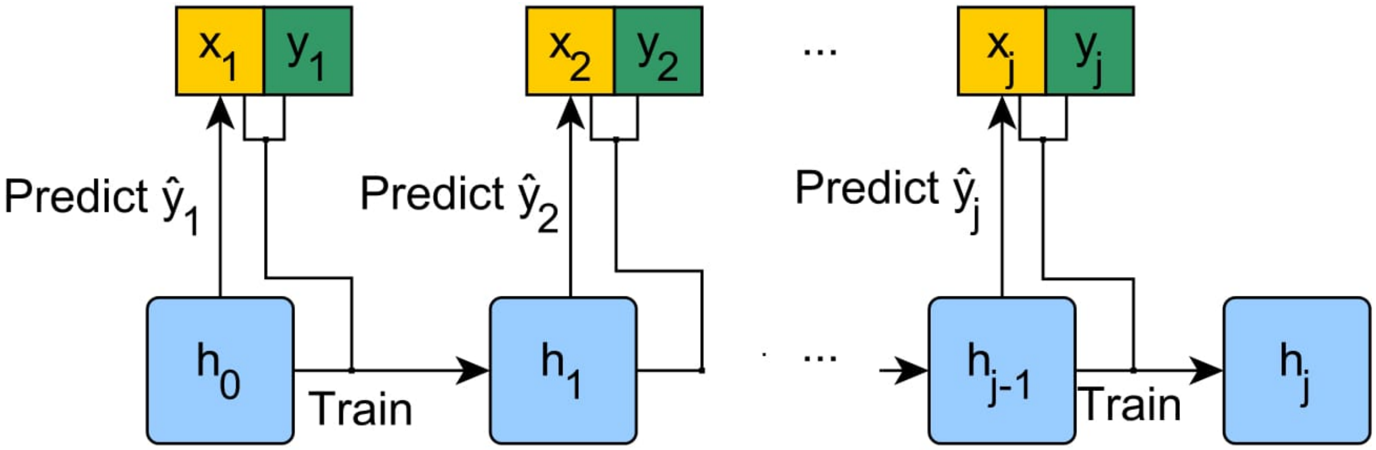
- On-line learning scheme
  - After each touple (x_i, y_i)generate a new model h_i to predict the next incoming example



Setting

- Supervised stream classification
  - Predict for an incoming stream of features $x_1, \ldots, x_j, x_i\ \mathbb{R}^n$ the corresponding labels $y_1, \ldots y_j, y_i \in \{1, \ldots, c\}$

- On-line learning scheme
  - After each touple (x_i, y_i)generate a new model h_i to predict the next incoming example



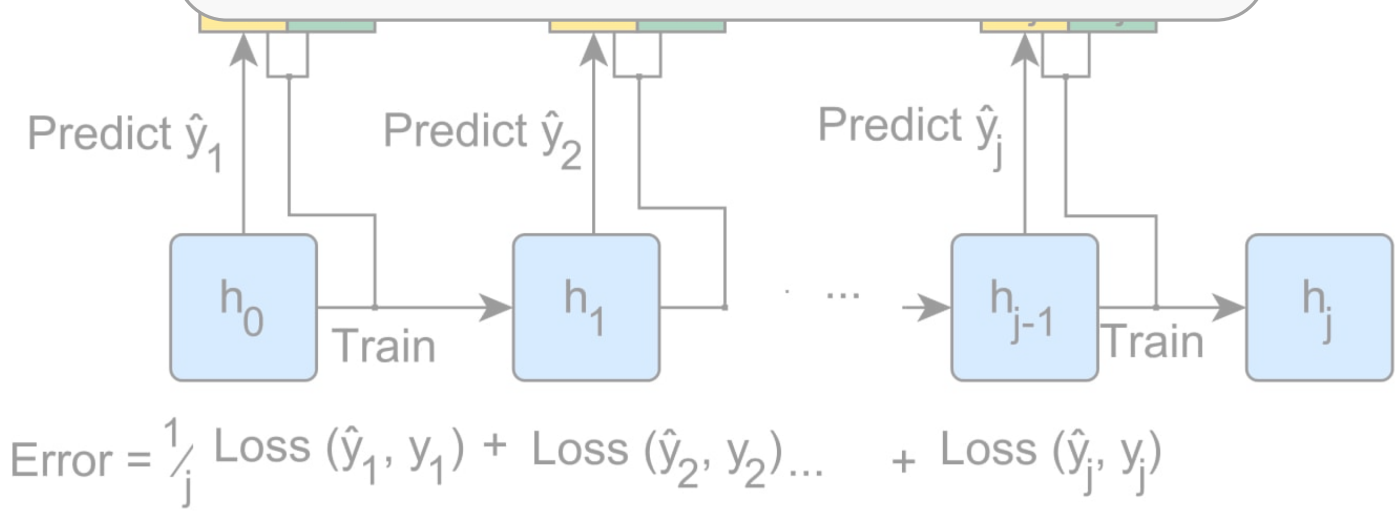
$$\text{Error} = \frac{1}{j}\ \text{Loss}(\hat{y}_1, y_1) + \text{Loss}(\hat{y}_2, y_2) \ldots + \text{Loss}(\hat{y}_j, y_j)$$

Setting

- Supervised stream classification
  - Predict for an incoming stream of features $x_1, \dots, x_j, x_i\ \mathbb{R}^n$ the corresponding labels $y_1, \dots y_j, y_i \in \{1, \dots, c\}$

- On-line lea
  - After each the next i

Preconditions for application:
- Obtainable labels in retrospective



Definition

- Concept drift is given when the joint distribution changes

$$\exists t_0, t_1 \colon P_{t_0}(X, Y) \neq P_{t_1}(X, Y)$$

Definition

➢ Concept drift is given when the joint distribution changes

$$\exists t_0, t_1 \colon P_{t_0}(X, Y) \neq P_{t_1}(X, Y)$$

t_0

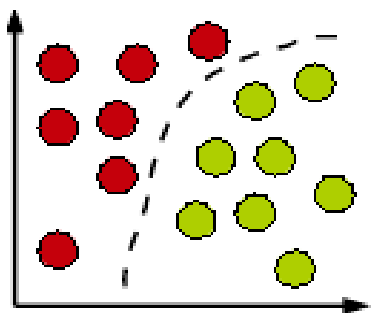


Image source : Gama et al. “A survey on concept drift adaptation”, ACM Computing Surveys 2014

Definition

- Concept drift is given when the joint distribution changes

$$\exists t_0, t_1 : P_{t_0}(X,Y) \neq P_{t_1}(X,Y)$$

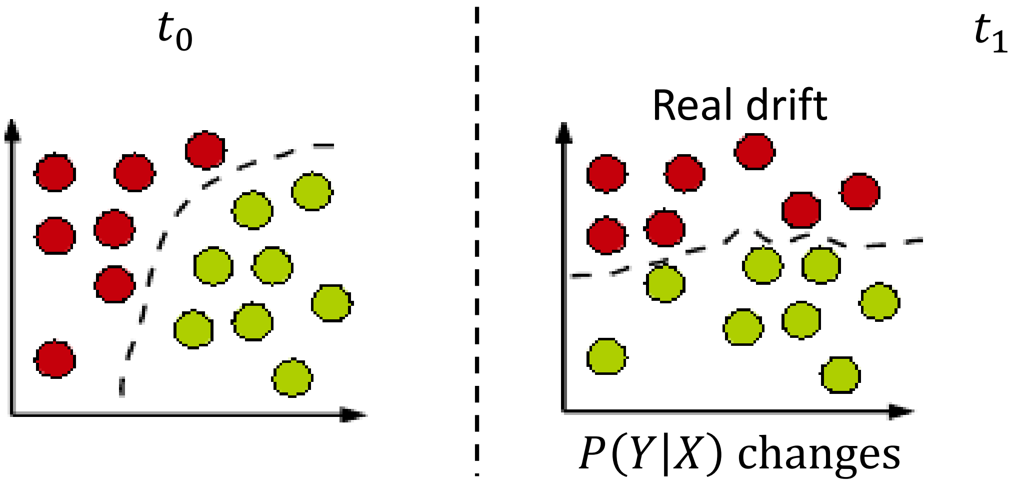


Image source : Gama et al. “A survey on concept drift adaptation”, ACM Computing Surveys 2014

Definition

- Concept drift is given when the joint distribution changes

$$\exists t_0, t_1 \colon P_{t_0}(X,Y) \neq P_{t_1}(X,Y)$$

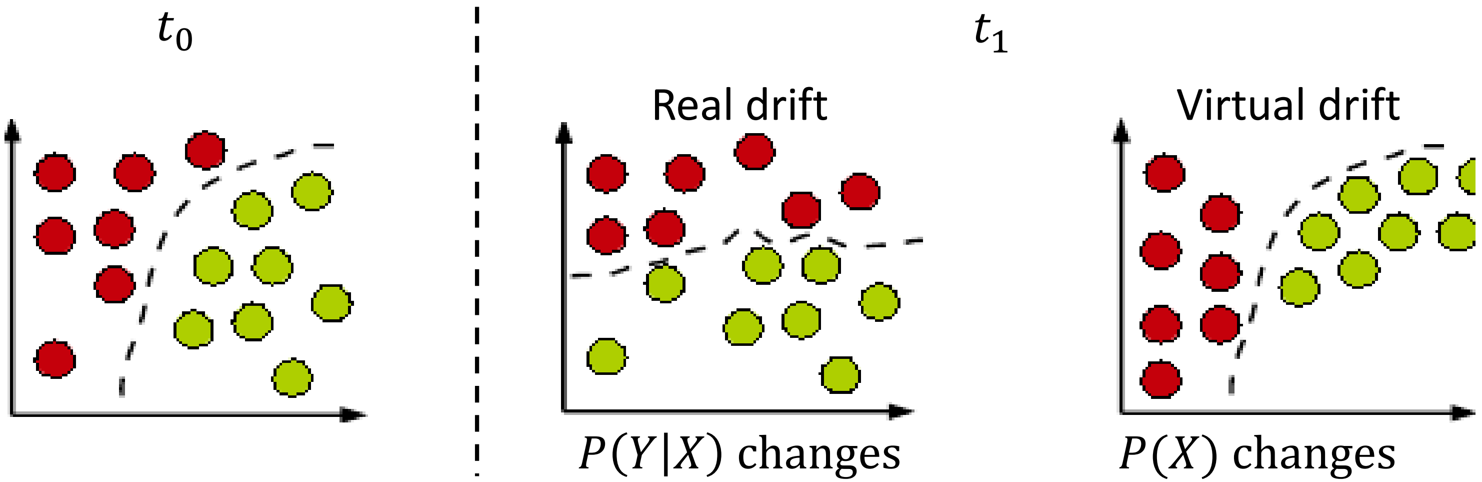
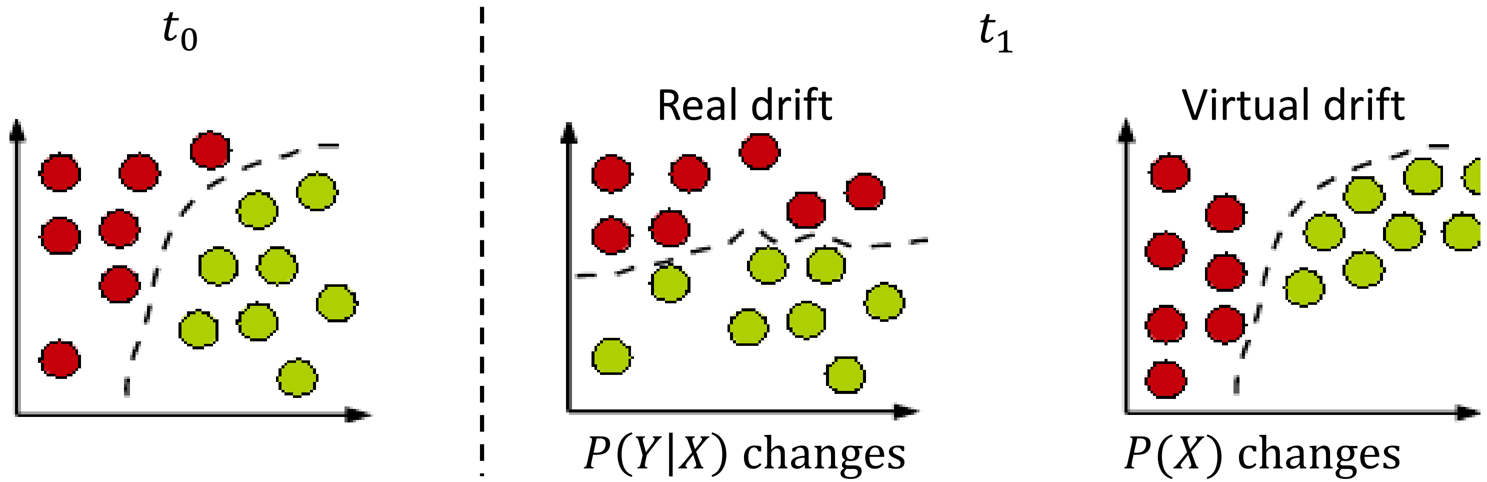


Image source : Gama et al. “A survey on concept drift adaptation”, ACM Computing Surveys 2014

Definition

- Concept drift is given when the joint distribution changes

$$\exists t_0, t_1\colon P_{t_0}(X,Y) \neq P_{t_1}(X,Y)$$



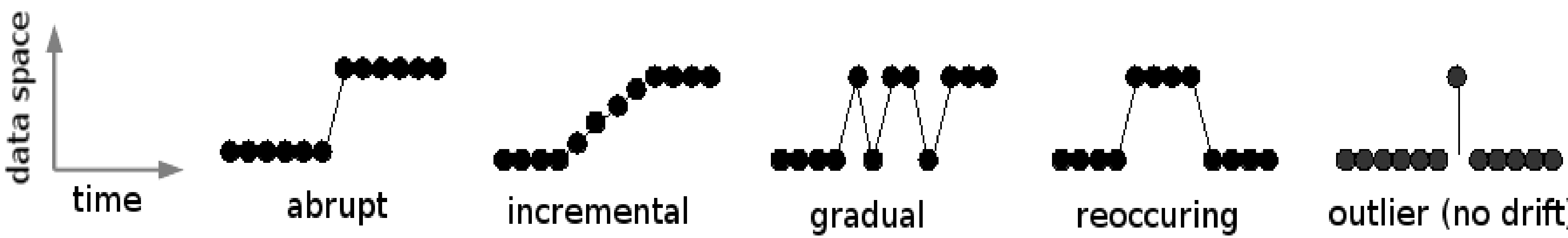


Image source : Gama et al. “A survey on concept drift adaptation”, ACM Computing Surveys 2014

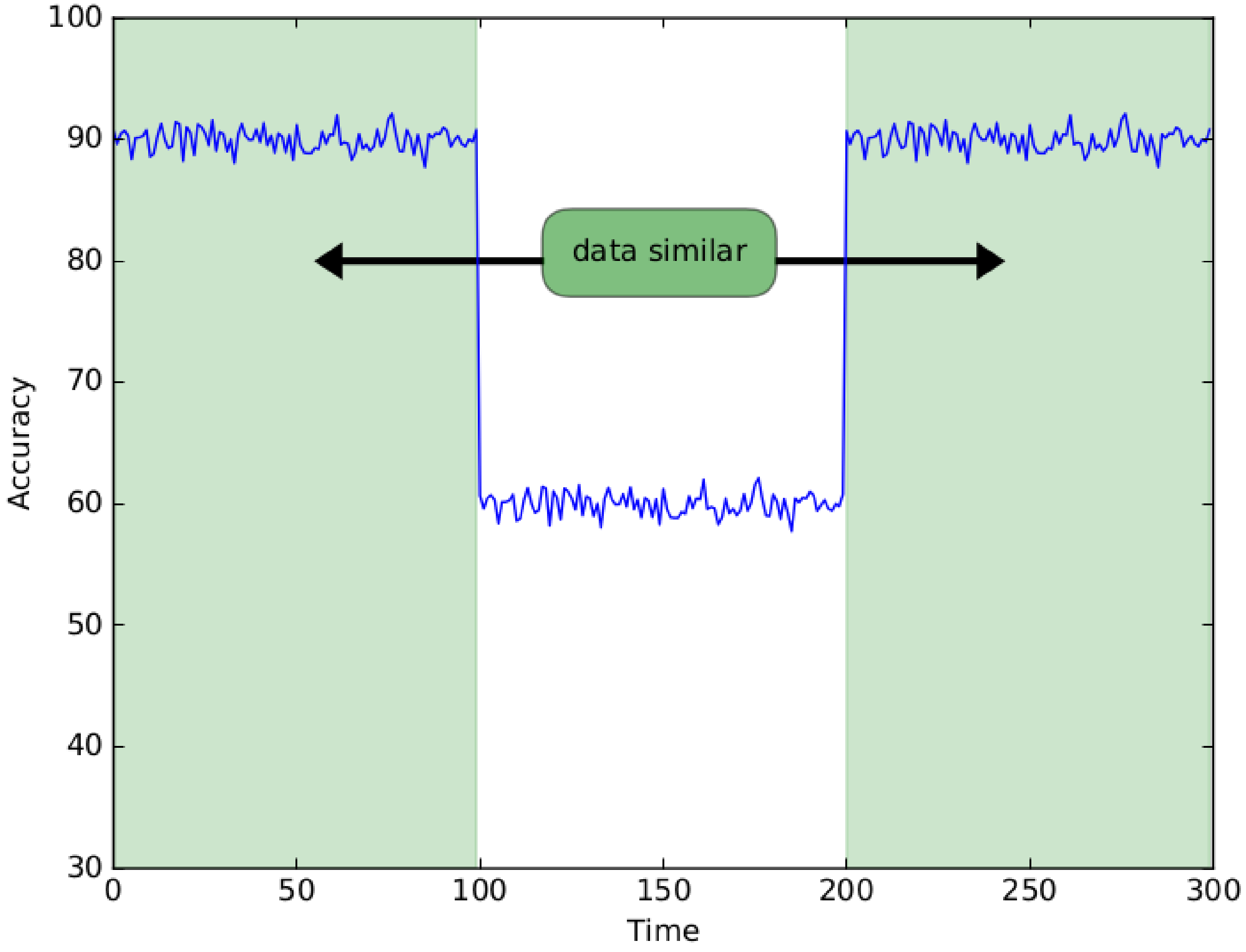
Related work

- Dynamic sliding windows techniques
  - PAW Bifet et al. “Efficient Data Stream Classification Via Probabilistic Adaptive Windows“, ACM 2013
- Ensemble methods with various weighting schemes
  - LVGB Bifet et al. “Leveraging Bagging for Evolving Data Streams“, ECML-PKDD 2010
  - Learn++.NSE Elwell et al. “Incremental Learning in Non-Stationary Environments“, IEEE-TNN 2011
  - DACC Jaber et al. “Online Learning: Searching for the Best Forgetting Strategy Under Concept Drift“, ICONIP-2013

Related work

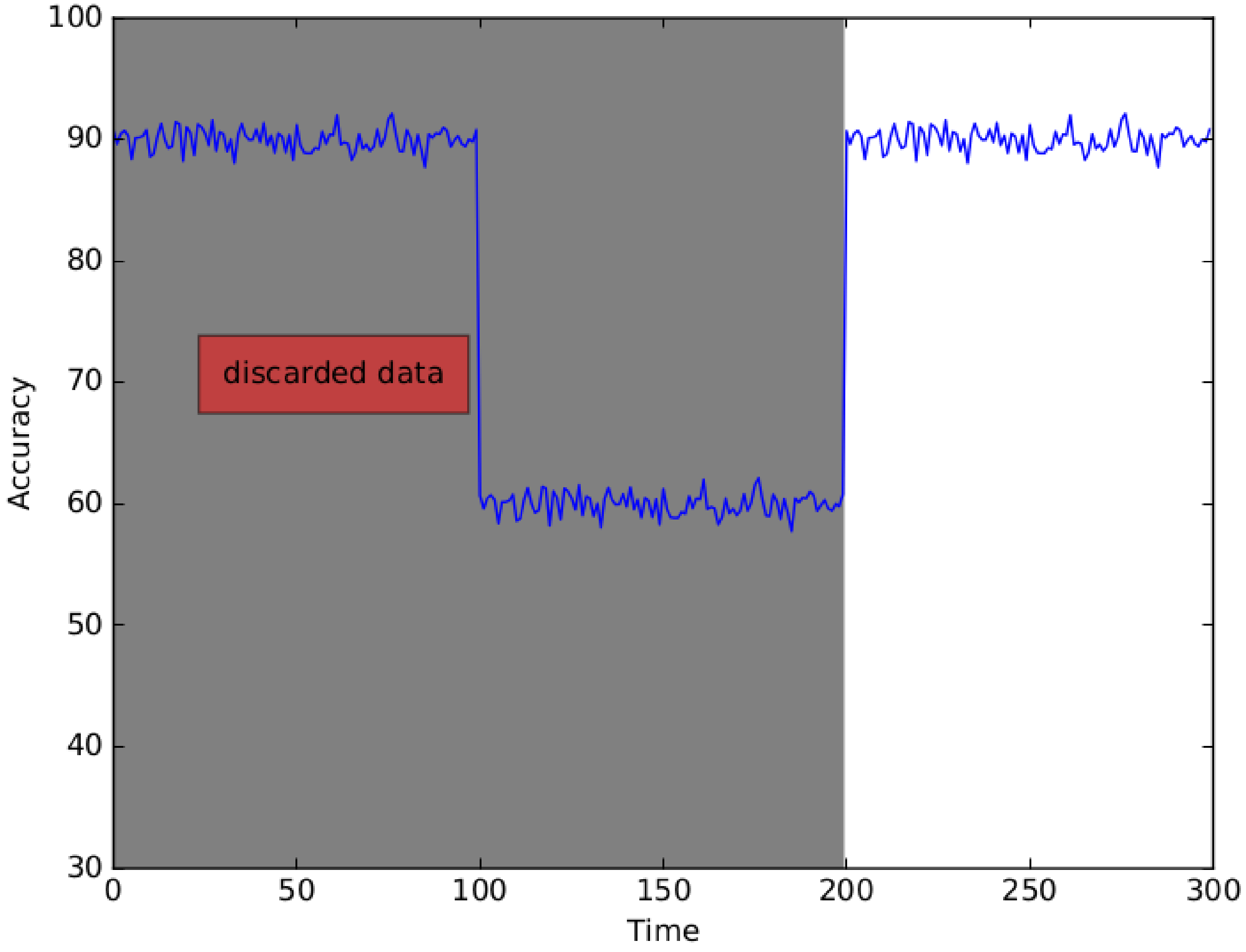
- Dynamic sliding windows techniques
  - PAW Bifet et al. “Efficient Data Stream Classification Via Probabilistic Adaptive Windows“, ACM 2013
- Ensemble methods with various weighting schemes
  - LVGB Bifet et al. “Leveraging Bagging for Evolving Data Streams“, ECML-PKDD 2010
  - Learn++.NSE Elwell et al. “Incremental Learning in Non-Stationary Environments“, IEEE-TNN 2011
  - DACC Jaber et al. “Online Learning: Searching for the Best Forgetting Strategy Under Concept Drift“, ICONIP-2013
- Drawbacks:
  - Target specific drift types
  - Require hyperparameter setting according to the expected drift
  - Discard former knowledge that still may be valuable

Drawbacks

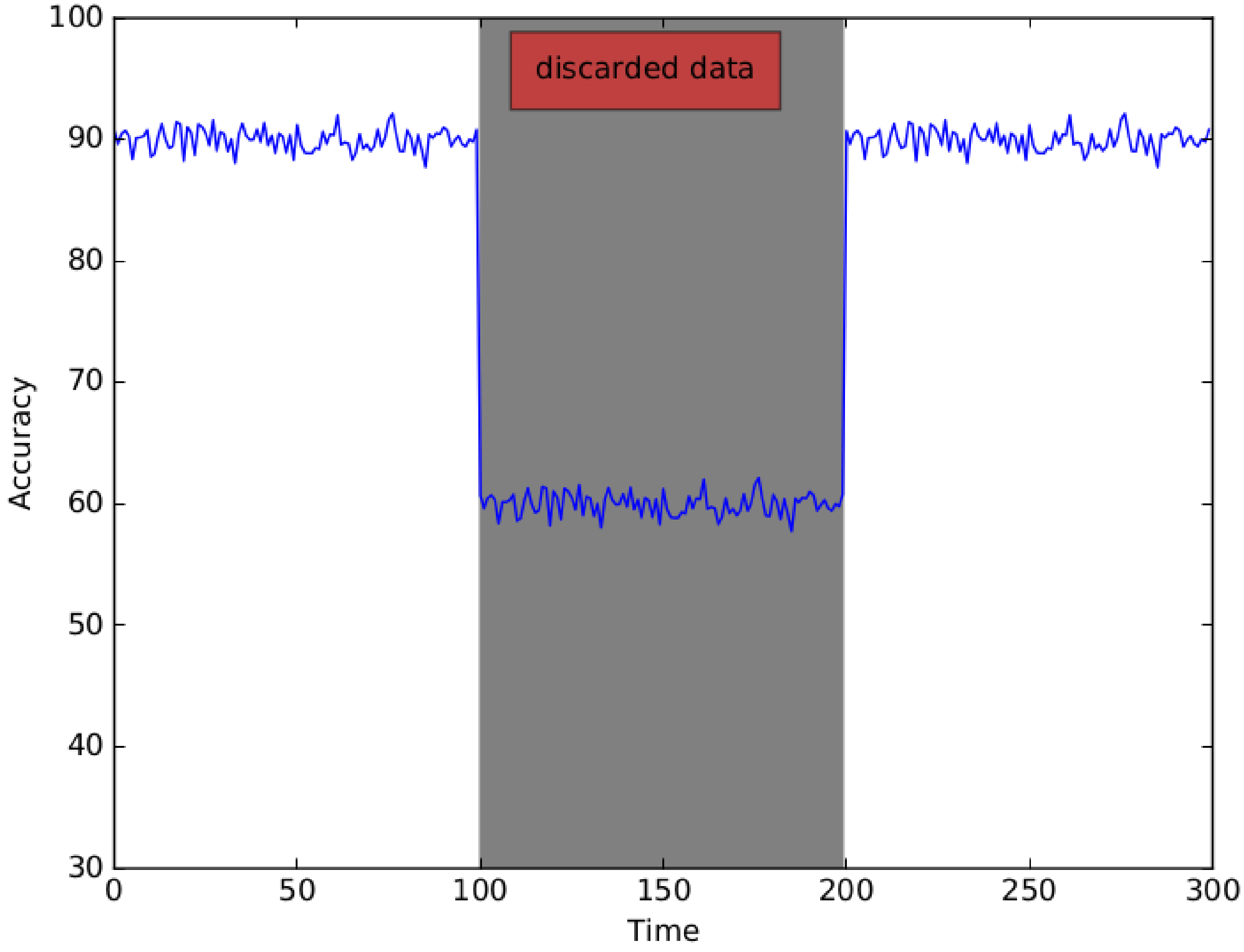


data similar
Accuracy
Time
100
90
80
70
60
50
40
30
0
50
100
150
200
250
300

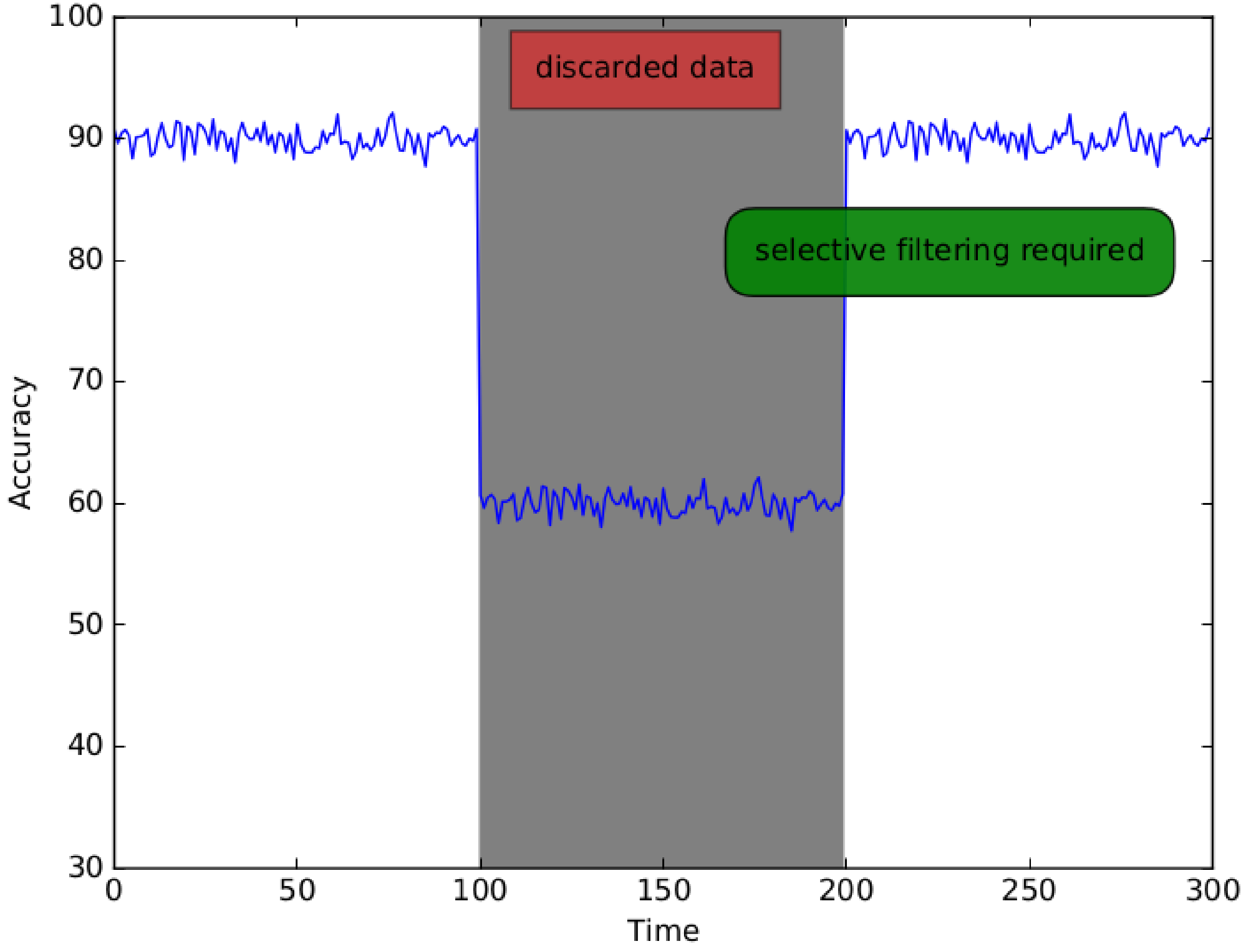
Drawbacks – Usual result



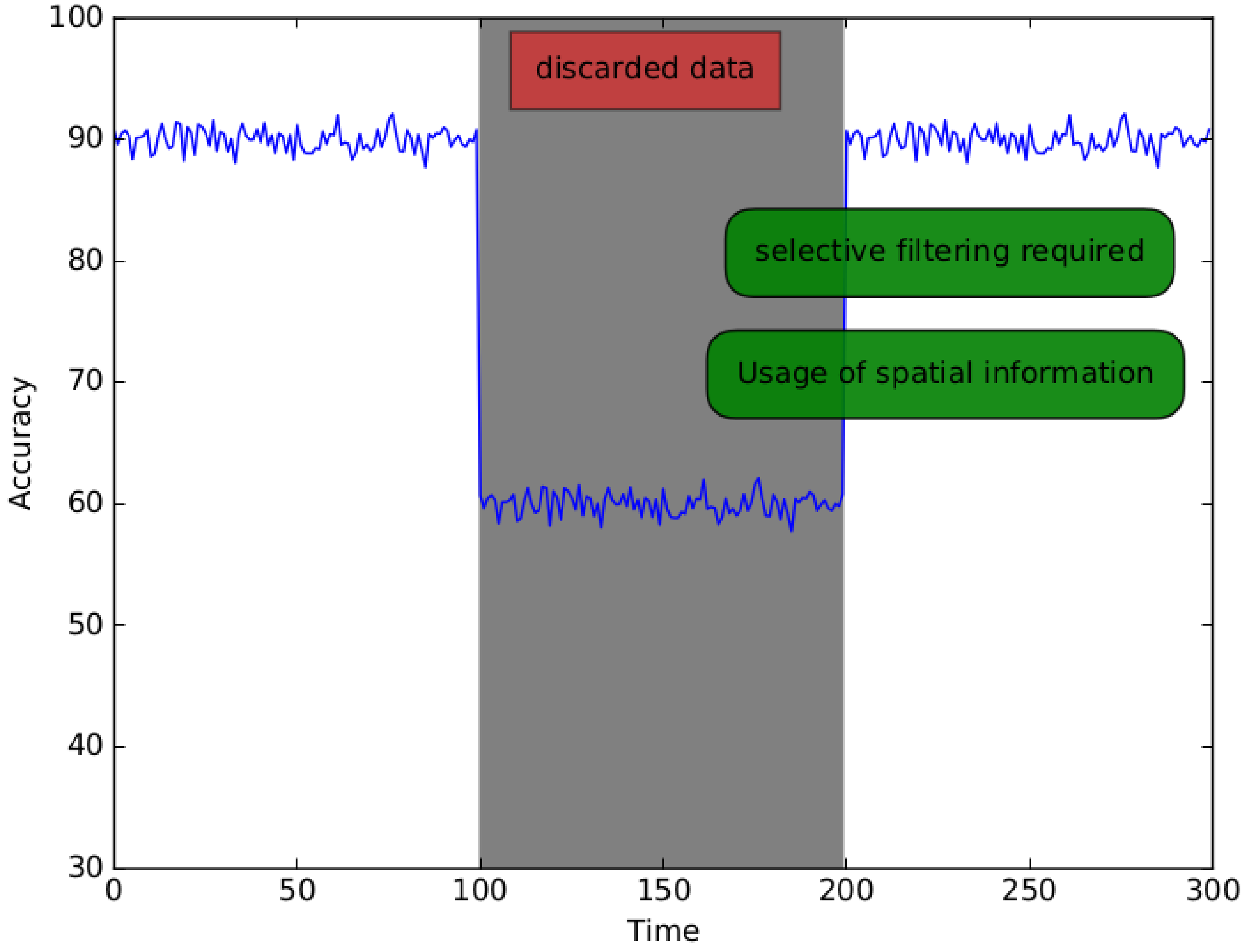
Drawbacks – Desired behavior



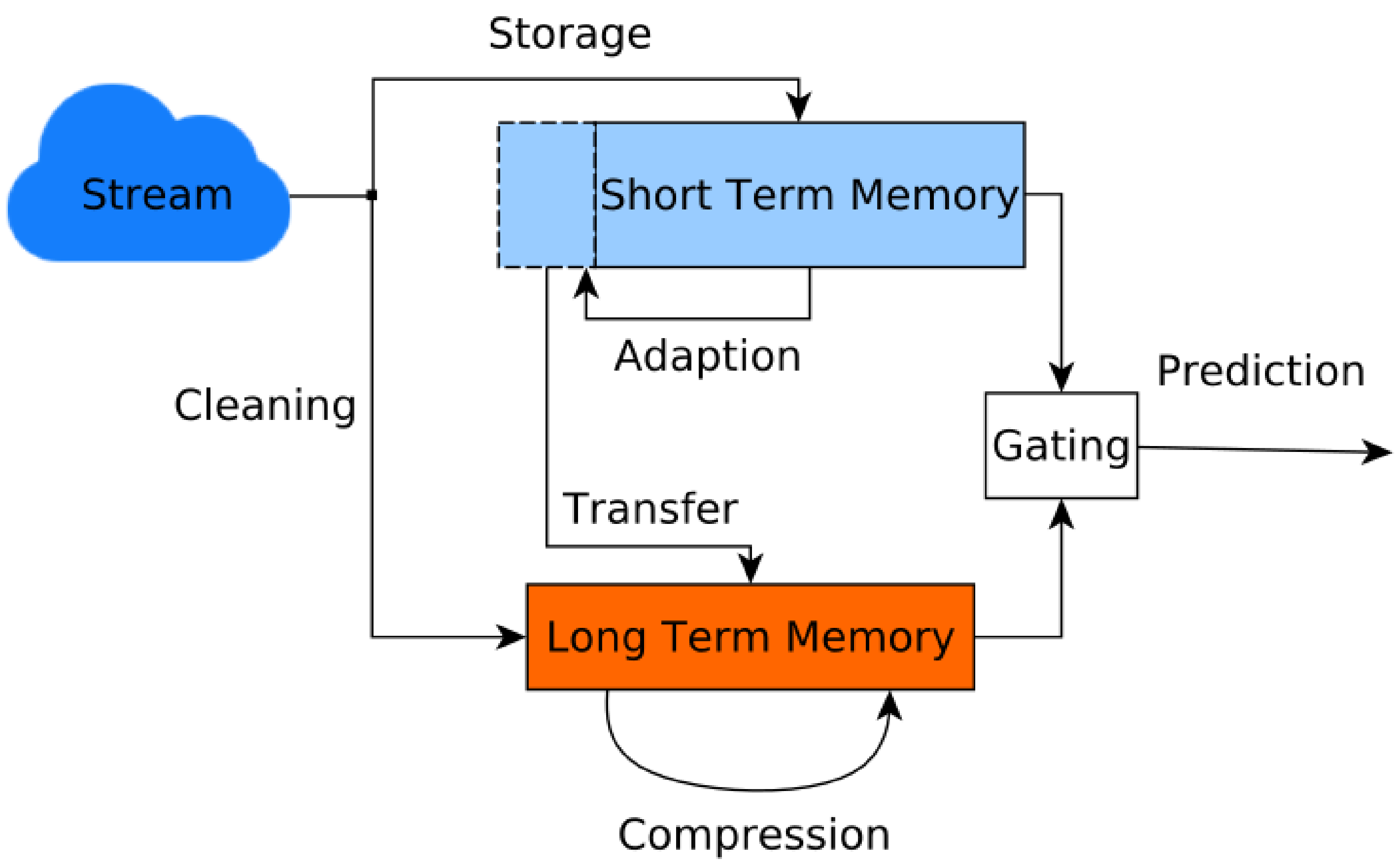
Drawbacks – Desired behavior



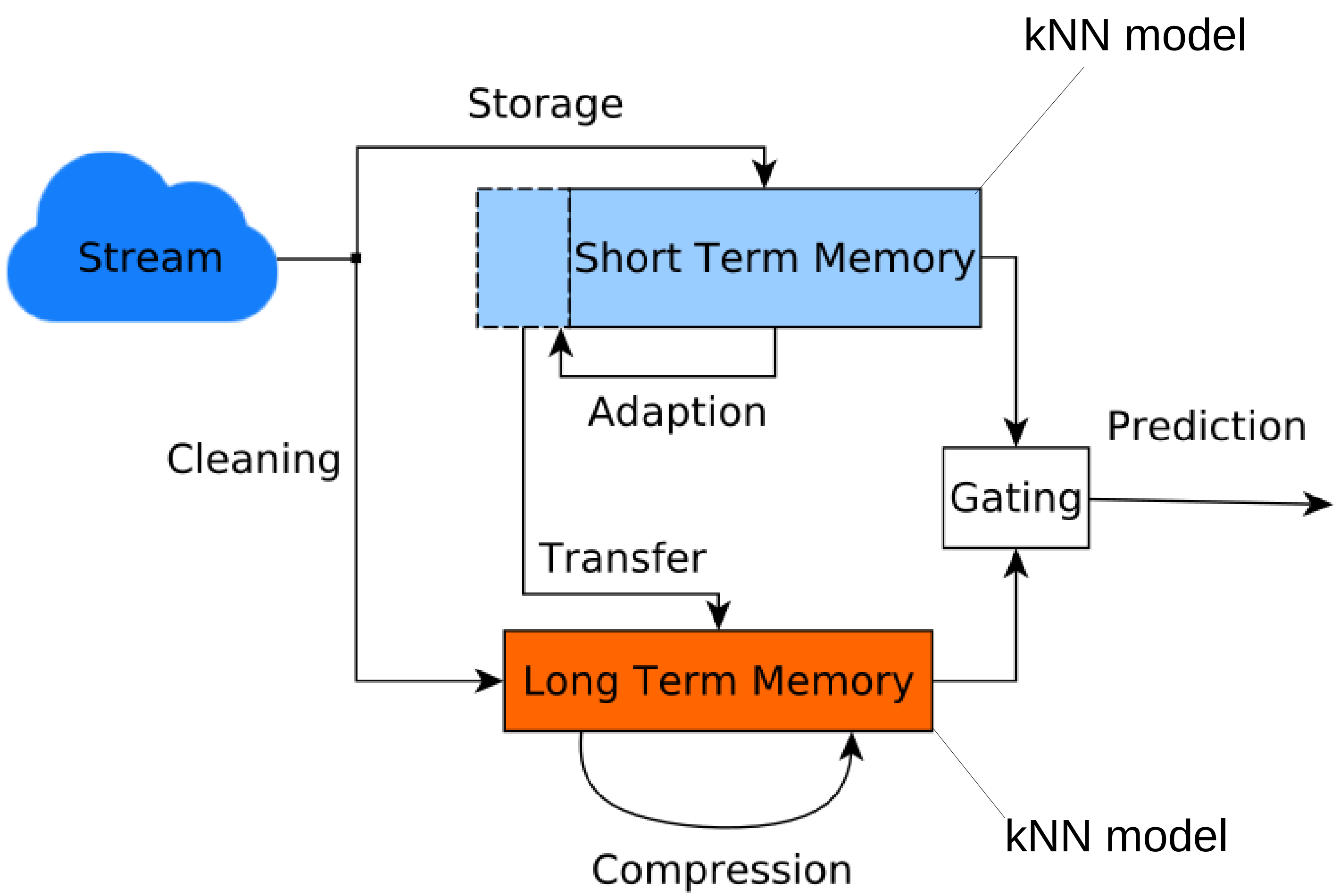
Drawbacks – Desired behavior



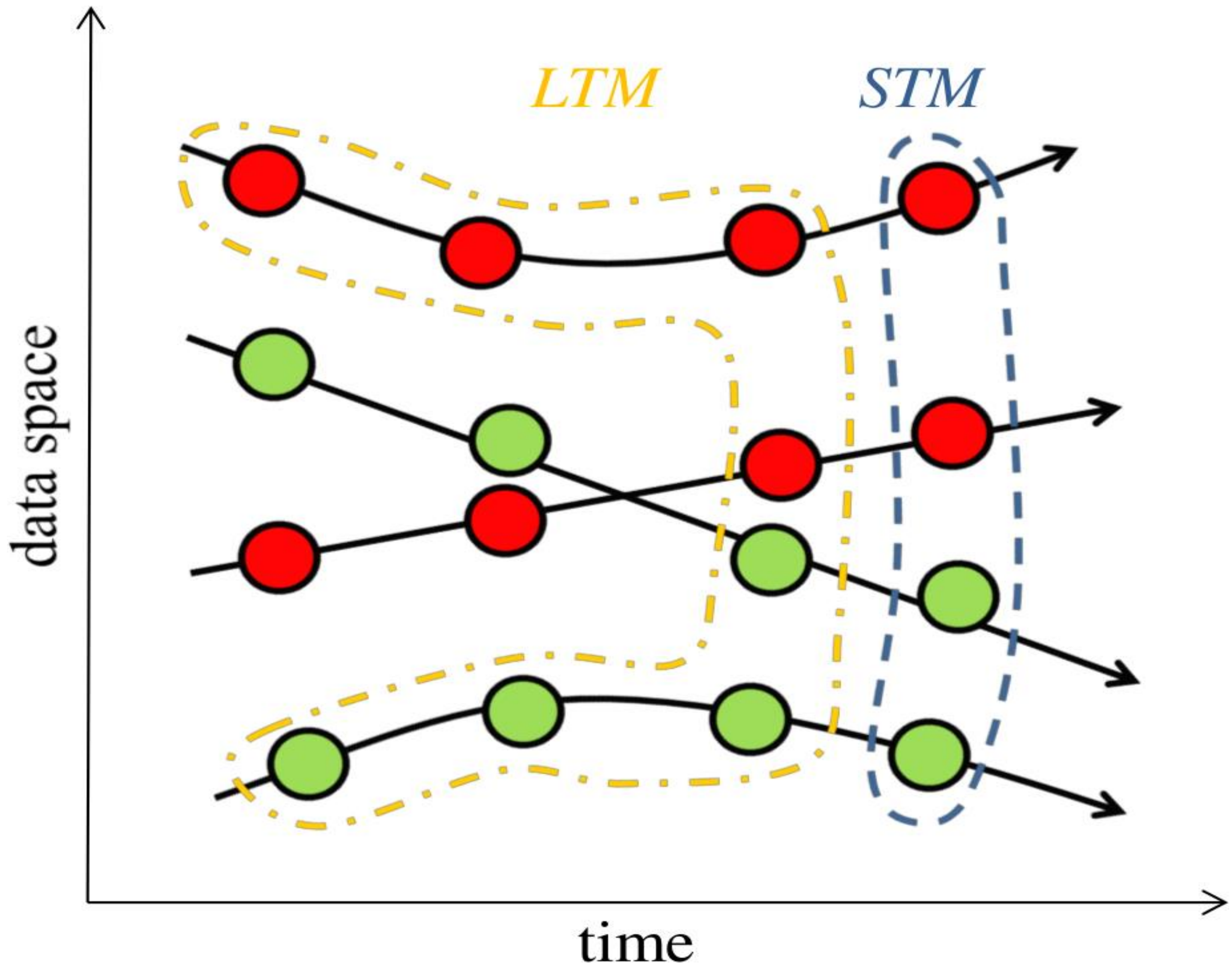
Self Adaptive Memory (SAM)



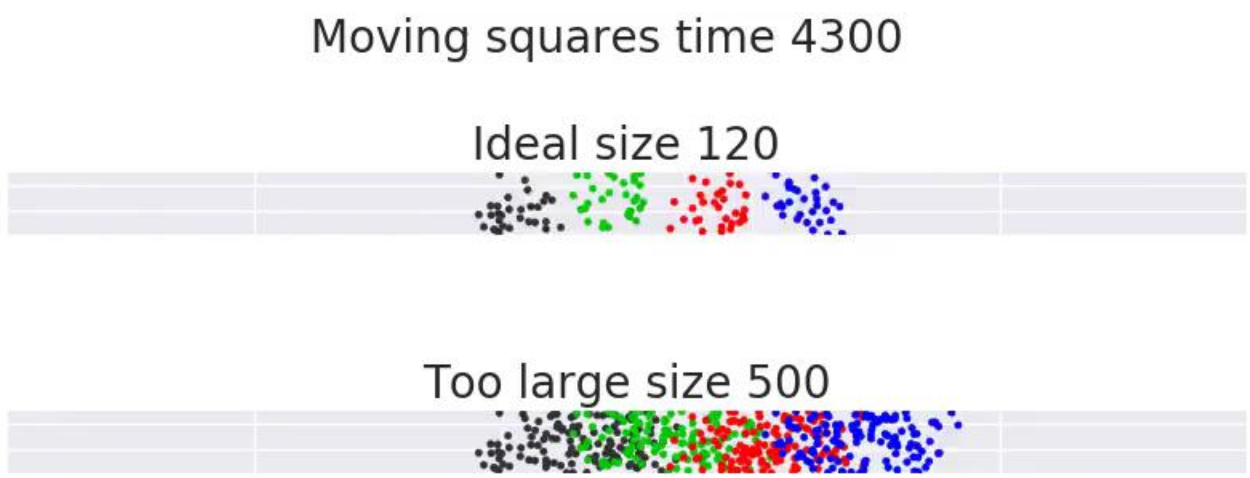
Self Adaptive Memory (SAM)



Self Adaptive Memory (SAM)



Moving squares dataset



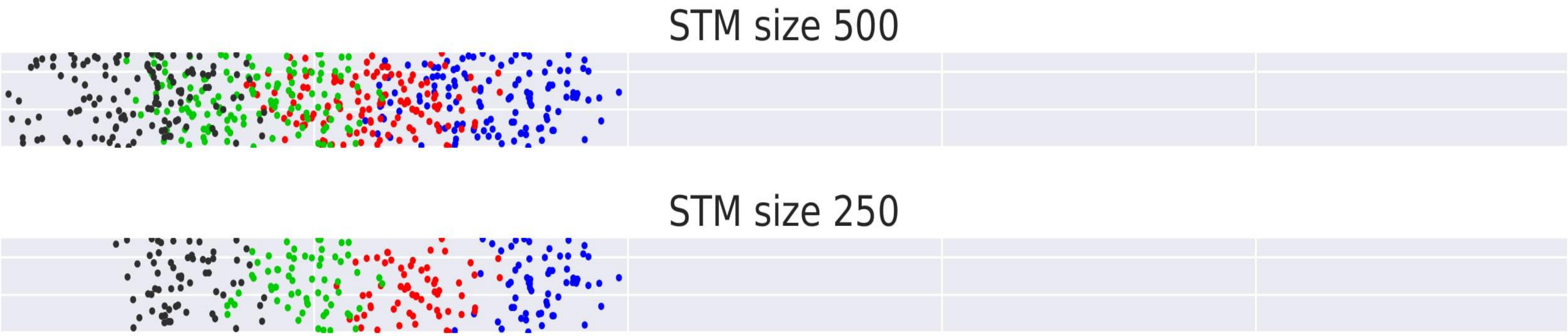
STM size adaptation

STM size adaptation

Error

27.12 %

STM size adaptation



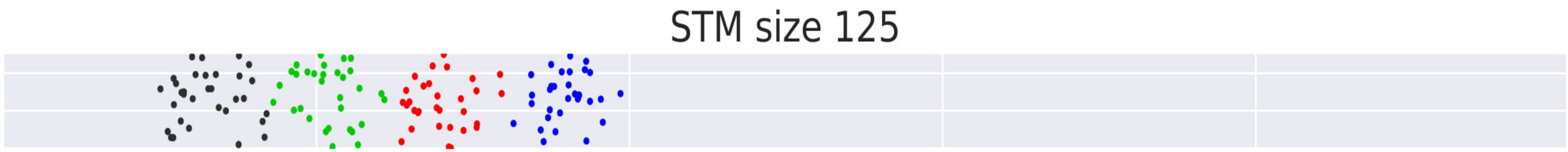
	Error
STM size 500	27.12 %
STM size 250	13.12 %

STM size adaptation

Error

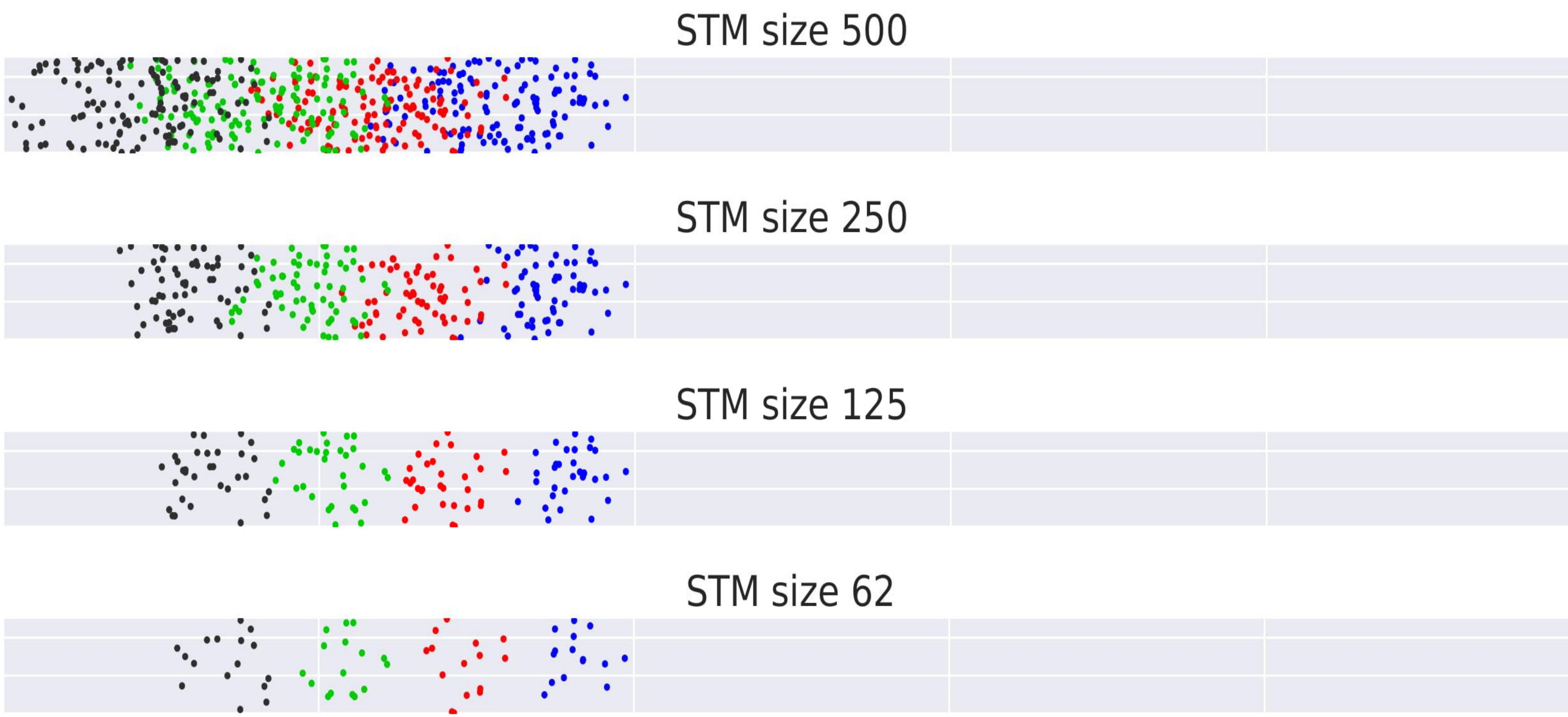
27.12 %

13.12 %



7.12 %

STM size adaptation



STM size	Error
500	27.12 %
250	13.12 %
125	7.12 %
62	0.0 %

STM size adaptation

Error

27.12 %

13.12 %

7.12 %

0.0 %

Distance-based cleaning

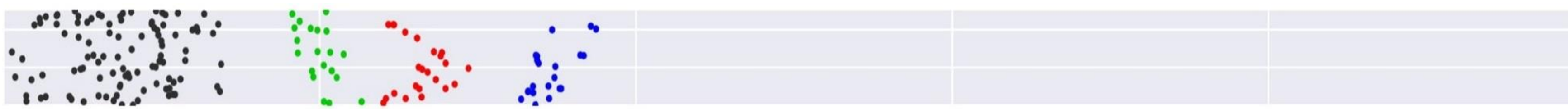
Distance-based cleaning

STM size 62

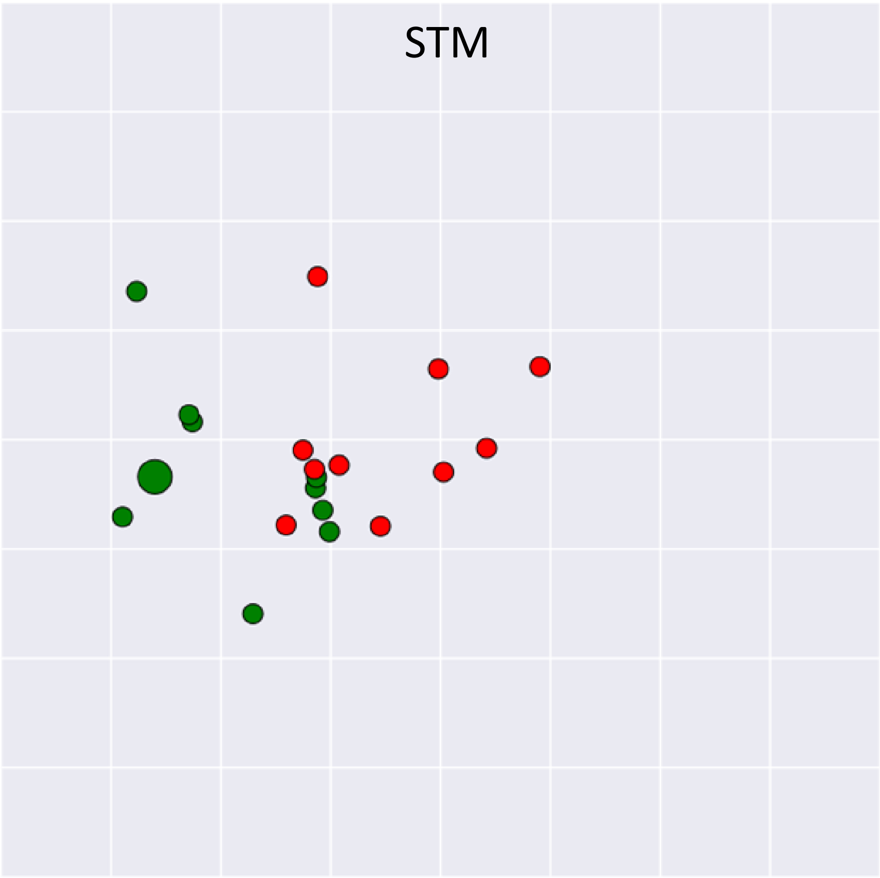
Dropped out data

cleaning

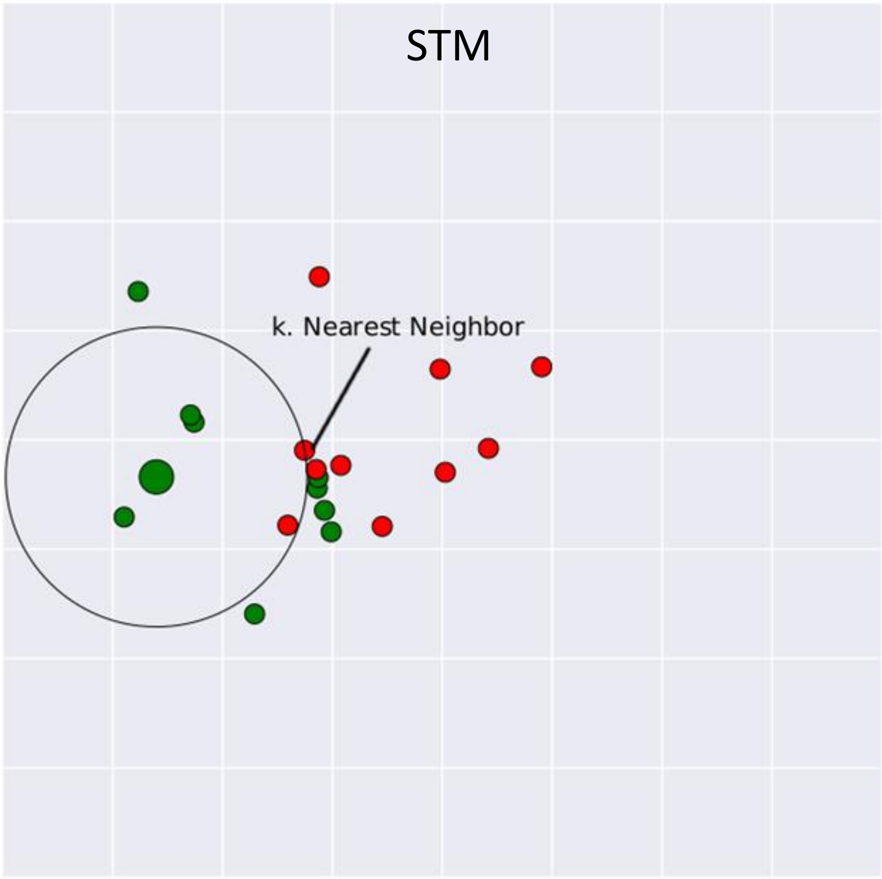
STM-consistent data



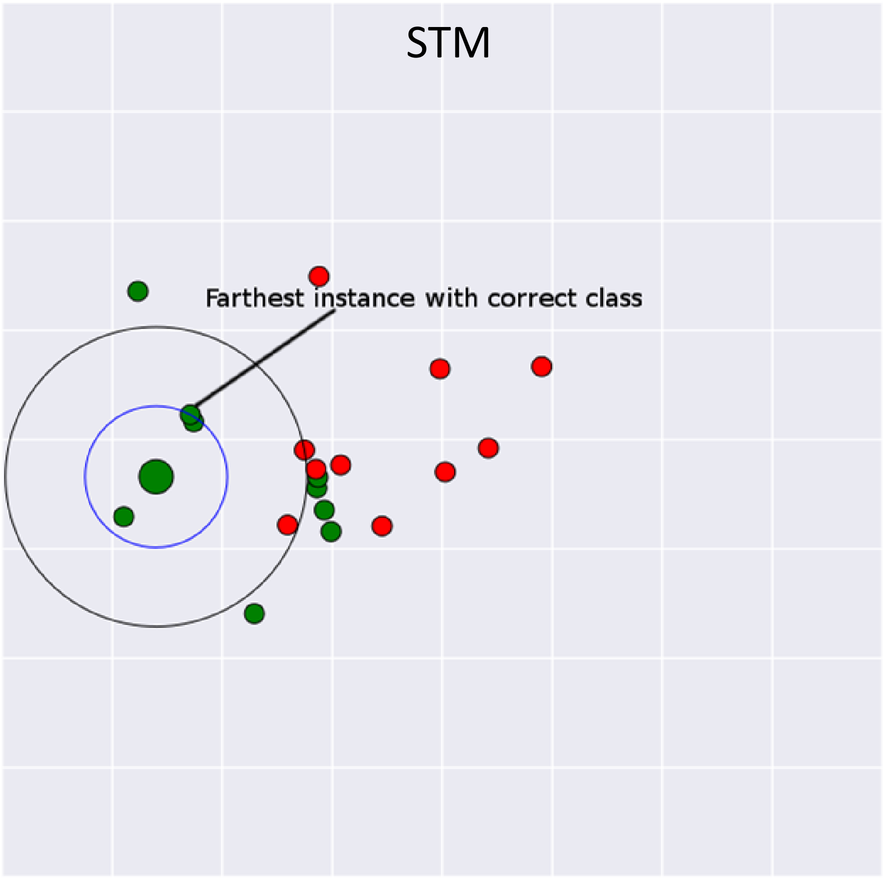
Distance-based cleaning



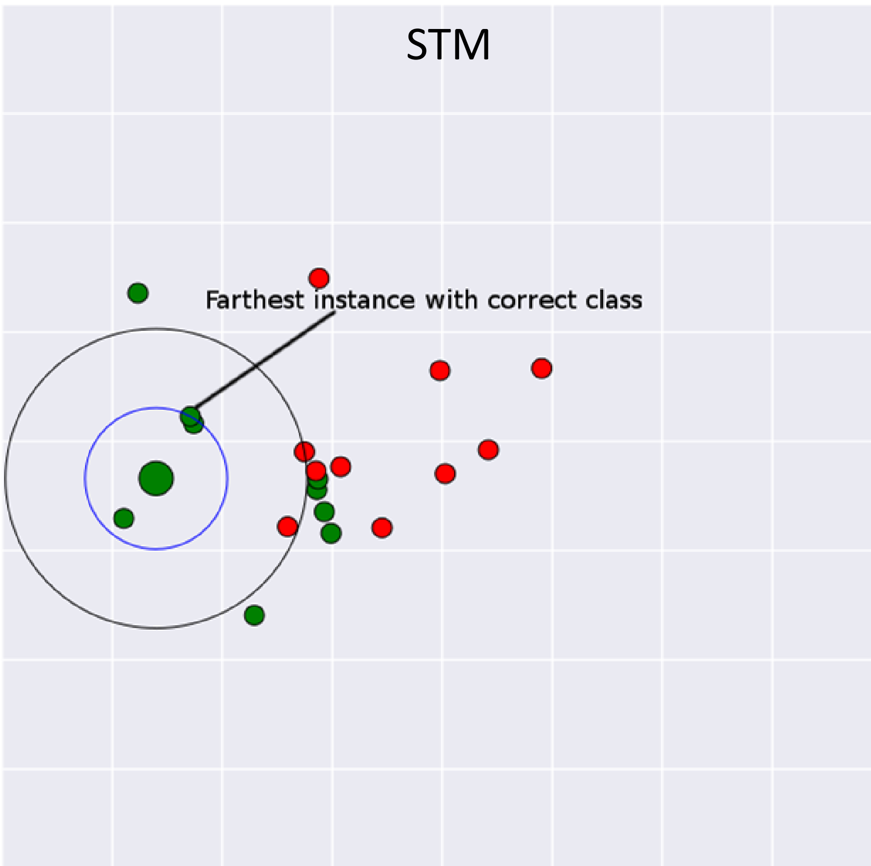
Distance-based cleaning

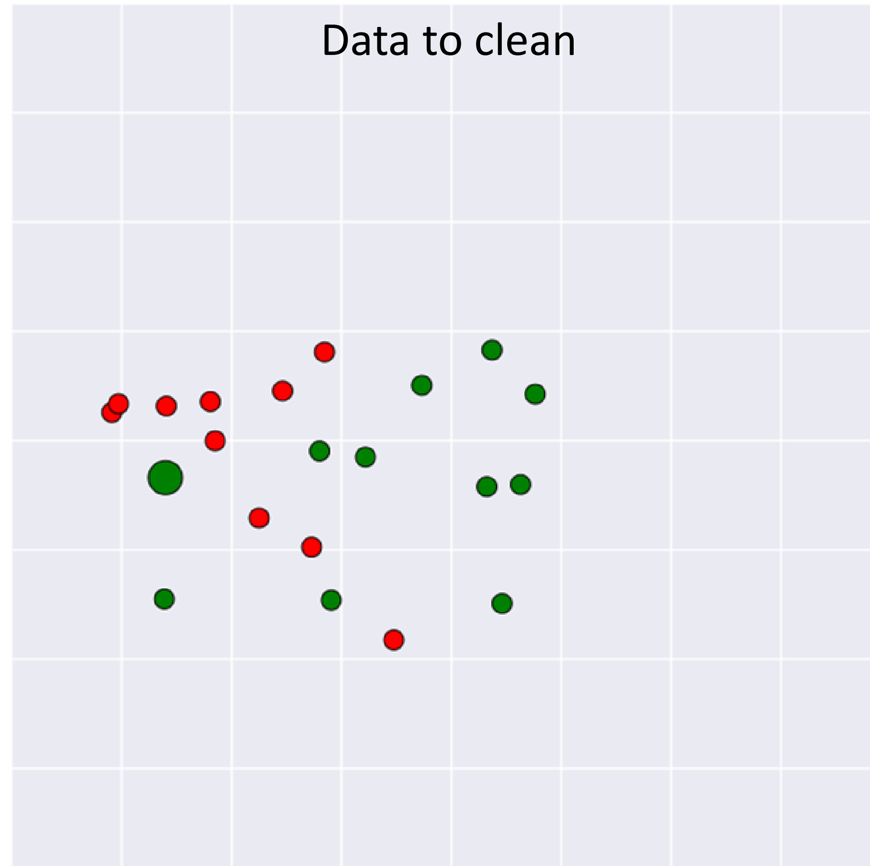


Distance-based cleaning

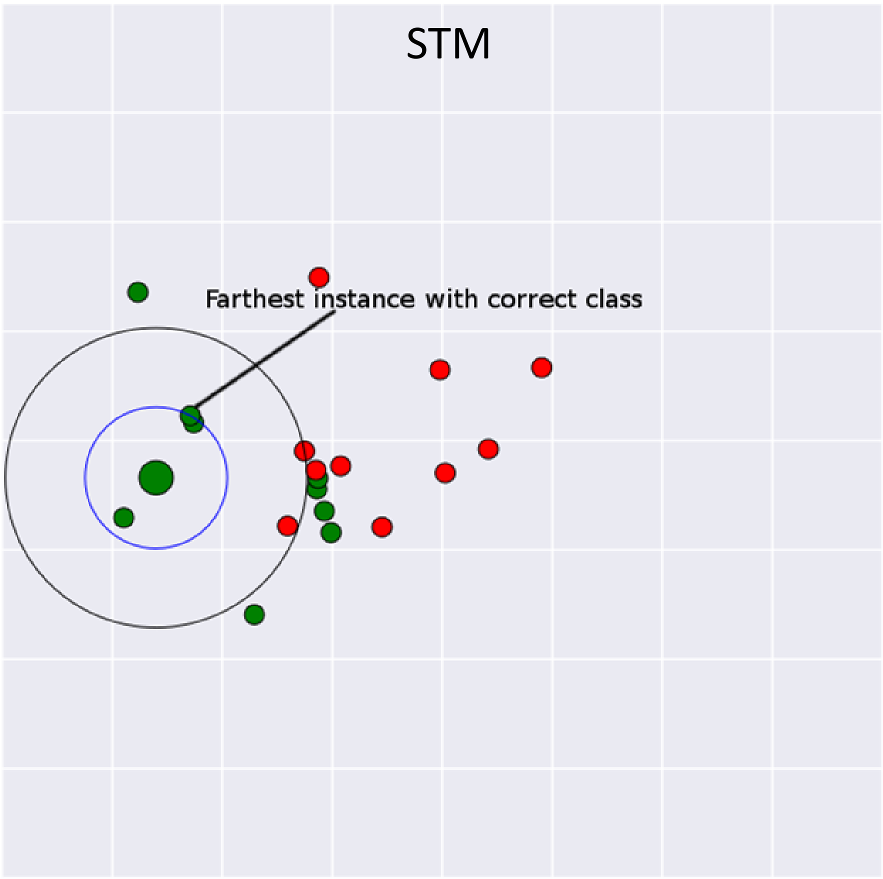


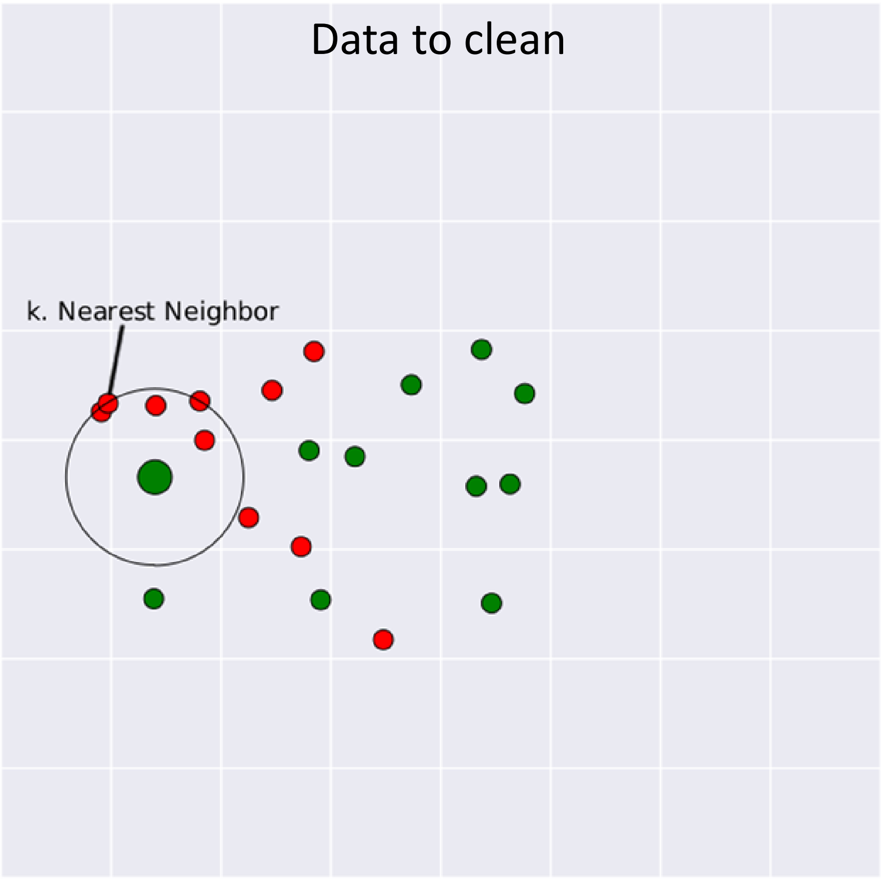
Distance-based cleaning



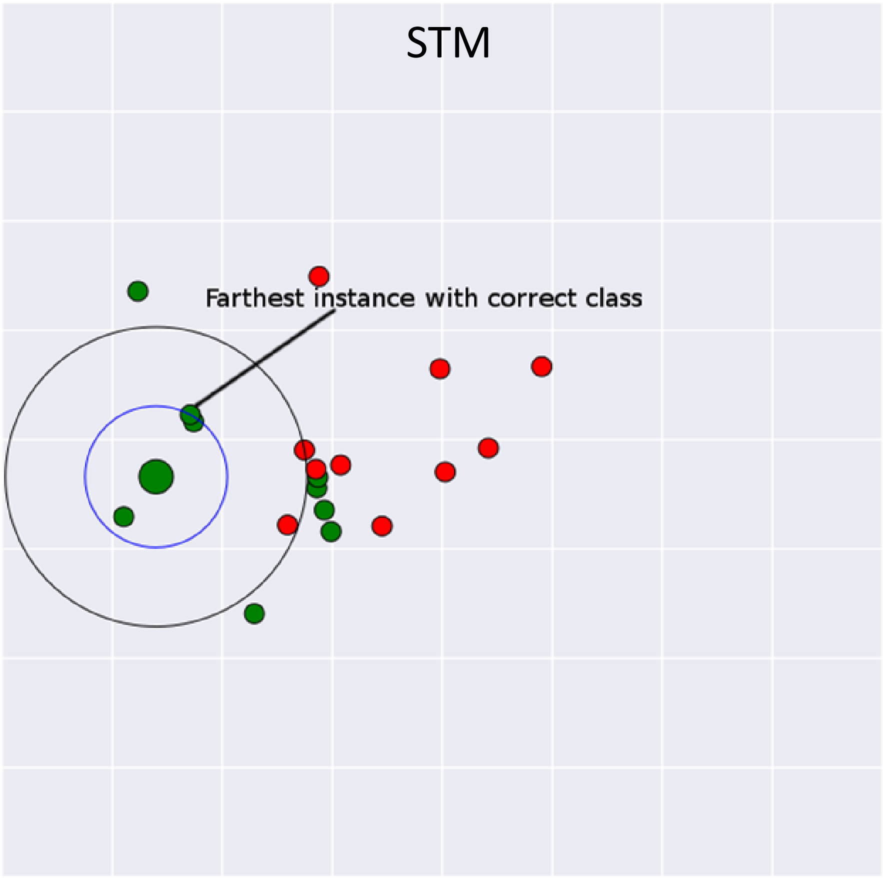


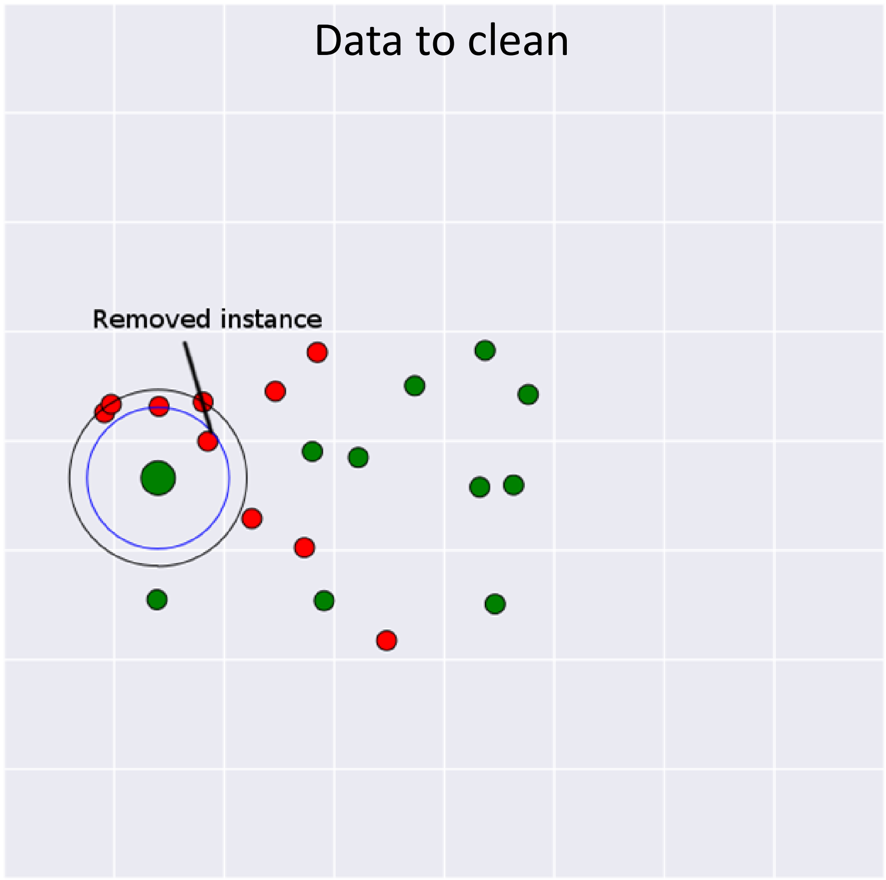
Distance-based cleaning



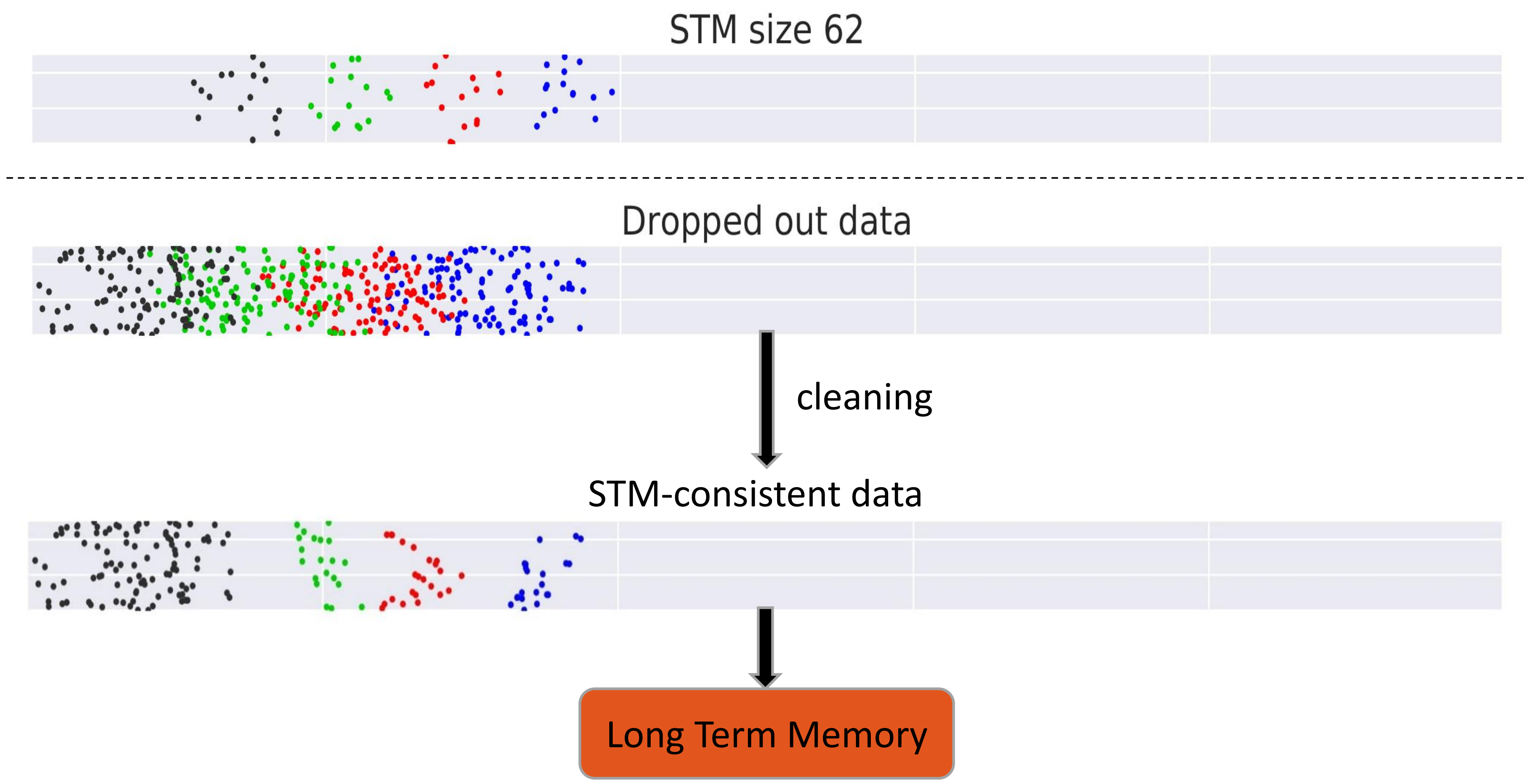


Distance-based cleaning

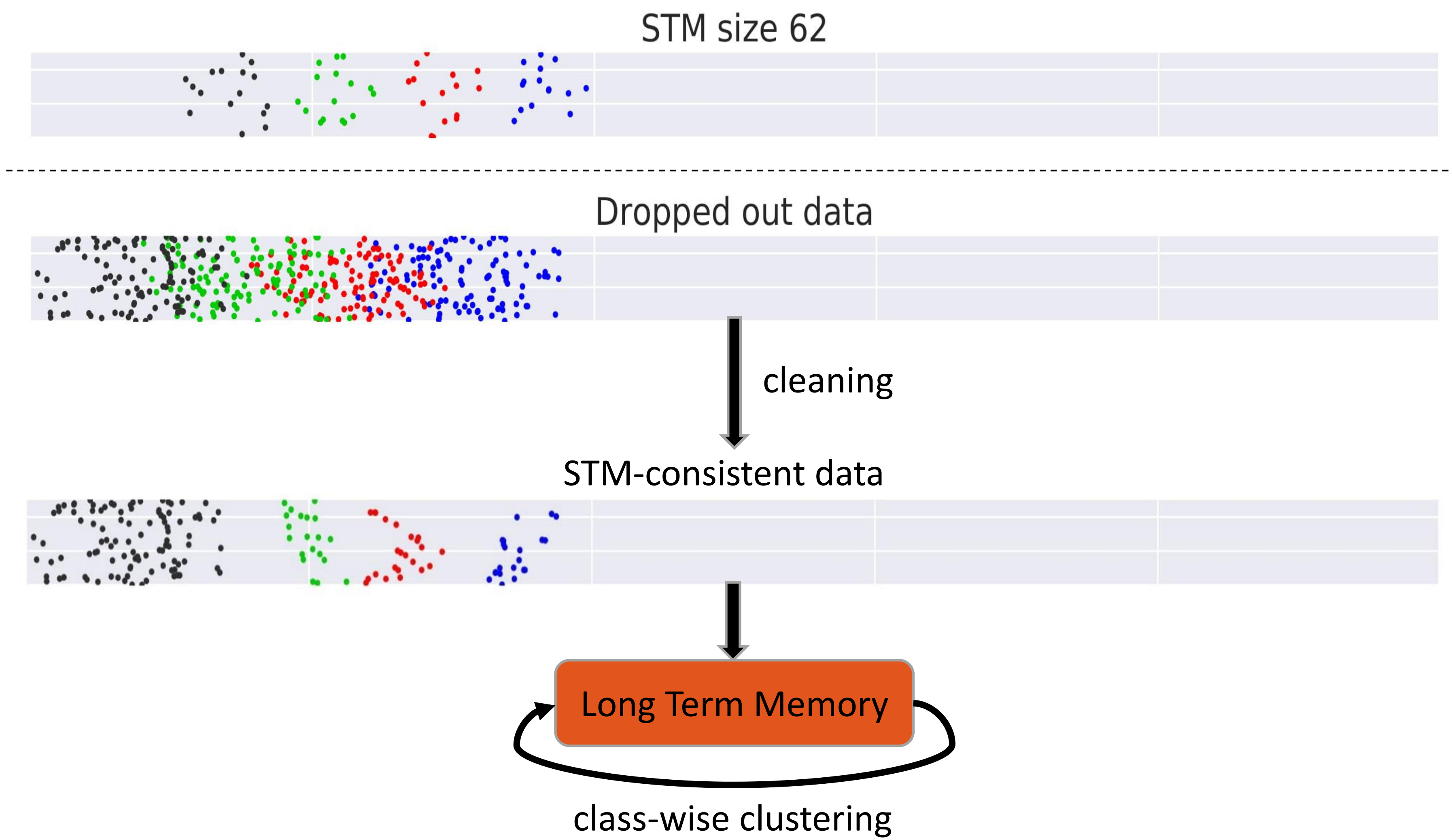




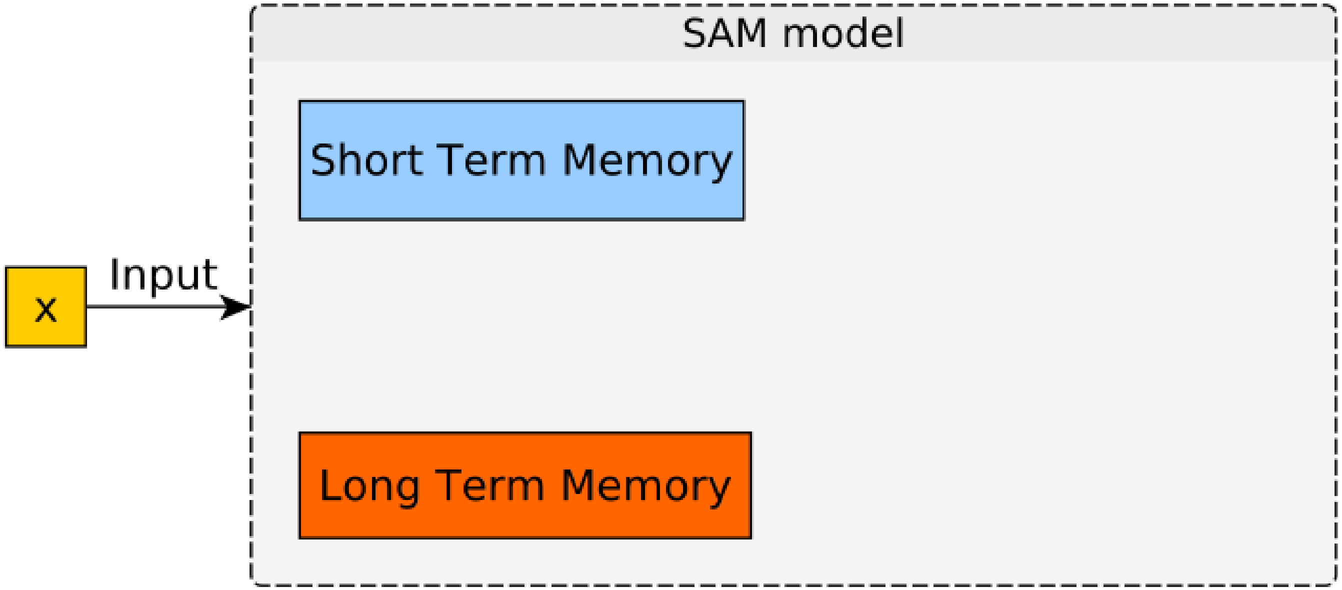
Adaptive compression



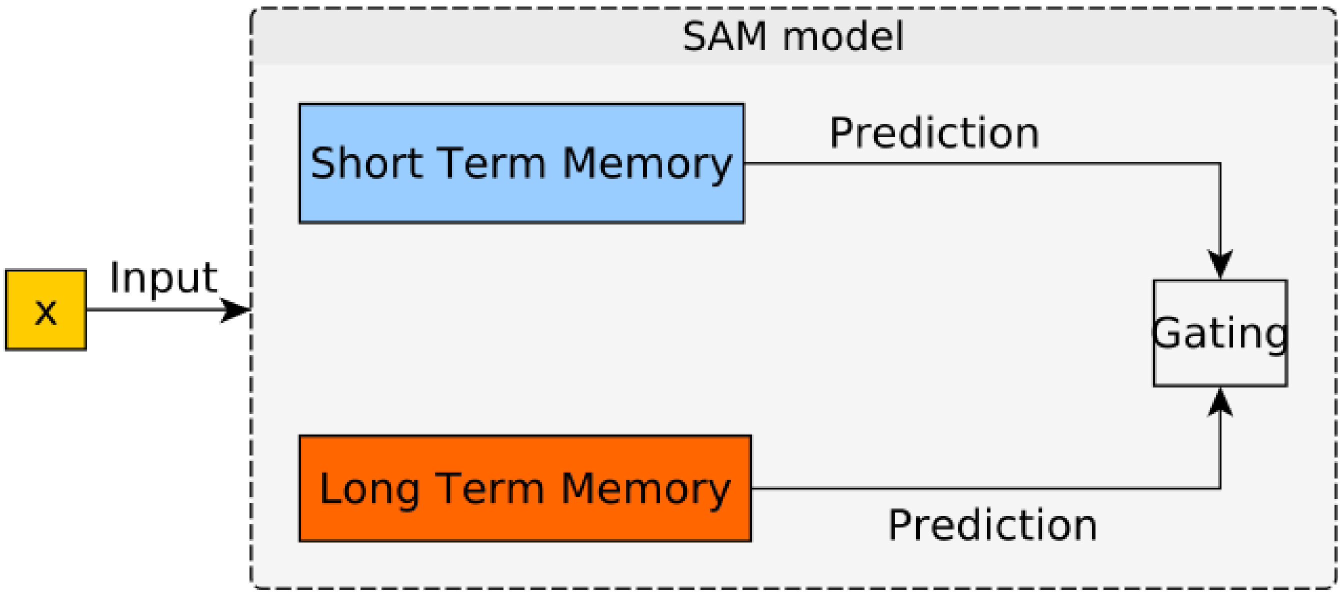
Adaptive compression



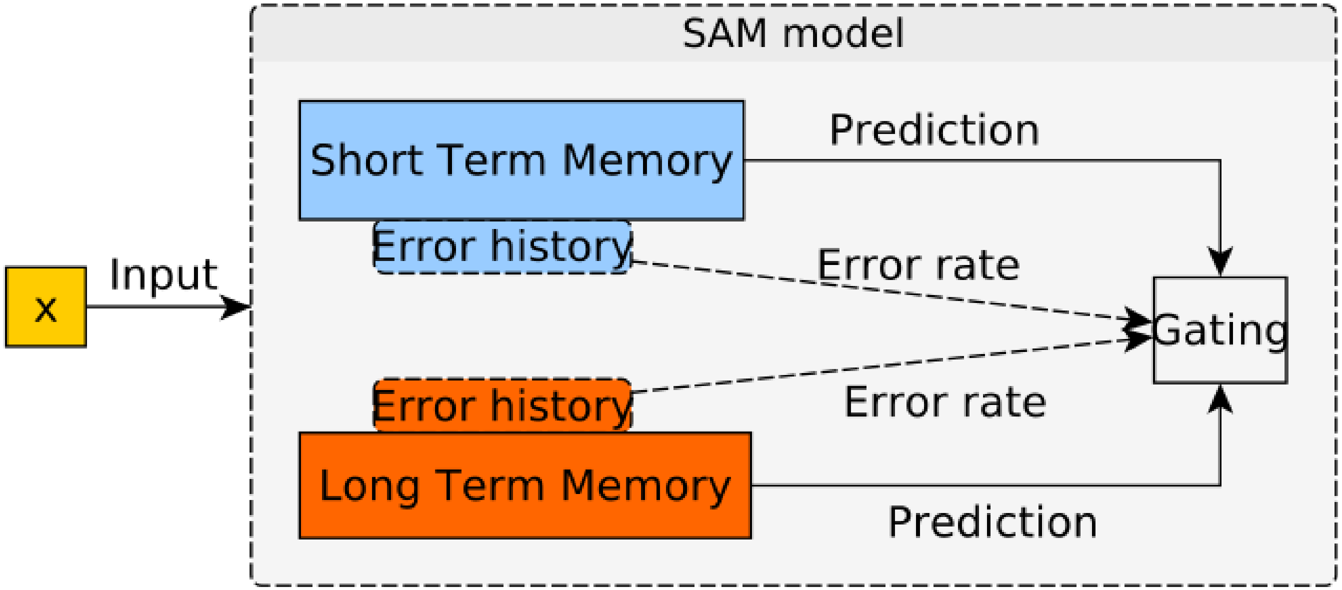
Prediction



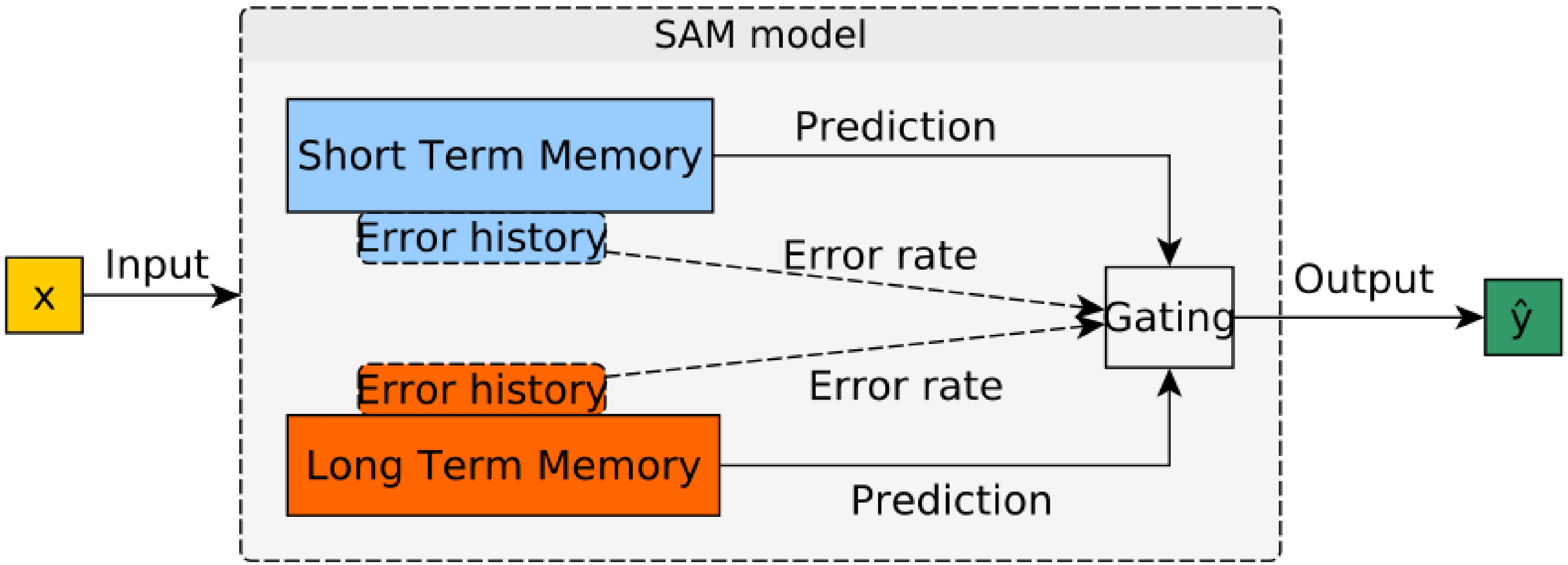
Prediction


SAM model
Short Term Memory
Prediction
x
Input
Gating
Long Term Memory
Prediction

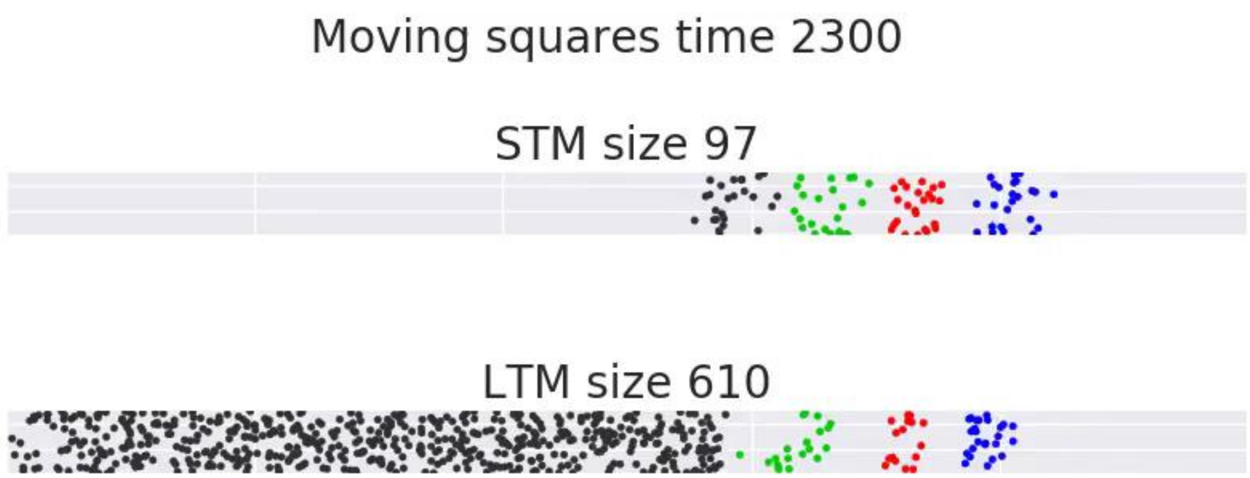
Prediction


x
Input
SAM model
Short Term Memory
Prediction
Error history
Error rate
Gating
Error history
Error rate
Long Term Memory
Prediction

Prediction


SAM model
Short Term Memory
Prediction
Error history
Error rate
x
Input
Gating
Output
ŷ
Error history
Error rate
Long Term Memory
Prediction

Moving squares by SAM



Results: Error rates / ranks

Dataset	LVGB	kNN_S	PAW	DACC	L++.NSE	**SAM**
SEA Concepts	**11.69**	13.83	13.39	15.68	14.48	12.50
Rotating Hyperplane	**12.53**	16.00	16.16	18.20	15.58	13.31
Moving RBF	44.84	20.36	24.04	54.34	44.50	**15.30**
Interchanging RBF	6.11	45.92	8.56	**1.40**	27.52	5.70
Moving Squares	12.17	68.87	61.01	**1.17**	65.90	2.30
Transient Chessb.	17.95	7.36	14.44	43.21	**1.98**	6.25
Mixed Drift	26.29	31.00	26.75	61.06	40.37	**13.33**
Artificial ∅	18.80	29.05	23.48	27.87	30.05	**9.81**
Artificial ∅ Rank	2.86	4.29	3.57	4.57	4.00	**1.71**
Weather	21.89	**21.53**	23.11	26.78	22.88	21.74
Electricity	**16.78**	28.61	26.13	16.87	27.24	17.52
Cover Type	9.07	**4.21**	6.76	10.05	15.00	4.8
Poker Hand	**13.65**	17.08	27.94	20.97	22.14	18.45
Outdoor	39.97	13.98	16.30	35.65	57.80	**11.25**
Rialto	39.64	22.74	24.96	28.93	40.36	**18.58**
Real world ∅	23.50	18.03	20.87	23.21	30.90	**15.40**
Real word ∅ Rank	3.17	2.33	4.00	4.17	5.33	**2.00**
Overall ∅	20.97	23.96	22.27	25.72	30.44	**12.39**
Overall ∅ Rank	3.00	3.38	3.77	4.38	4.62	**1.85**

SAM achieves best results

Dataset	LVGB	kNN$_S$	PAW	DACC	L++.NSE	**SAM**
SEA Concepts	**11.69**	13.83	13.39	15.68	14.48	12.50
Rotating Hyperplane	**12.53**	16.00	16.16	18.20	15.58	13.31
Moving RBF	44.84	20.36	24.04	54.34	44.50	**15.30**
Interchanging RBF	6.11	45.92	8.56	**1.40**	27.52	5.70
Moving Squares	12.17	68.87	61.01	**1.17**	65.90	2.30
Transient Chessb.	17.95	7.36	14.44	43.21	**1.98**	6.25
Mixed Drift	26.29	31.00	26.75	61.06	40.37	**13.33**
Artificial ∅	18.80	29.05	23.48	27.87	30.05	**9.81**
Artificial ∅ Rank	2.86	4.29	3.57	4.57	4.00	**1.71**
Weather	21.89	**21.53**	23.11	26.78	22.88	21.74
Electricity	**16.78**	28.61	26.13	16.87	27.24	17.52
Cover Type	9.07	**4.21**	6.76	10.05	15.00	4.8
Poker Hand	**13.65**	17.08	27.94	20.97	22.14	18.45
Outdoor	39.97	13.98	16.30	35.65	57.80	**11.25**
Rialto	39.64	22.74	24.96	28.93	40.36	**18.58**
Real world ∅	23.50	18.03	20.87	23.21	30.90	**15.40**
Real word ∅ Rank	3.17	2.33	4.00	4.17	5.33	**2.00**
Overall ∅	$\boxed{20.97}$	23.96	22.27	25.72	30.44	$\boxed{\textbf{12.39}}$
Overall ∅ Rank	$\boxed{3.00}$	3.38	3.77	4.38	4.62	$\boxed{\textbf{1.85}}$

SAM is robust

Dataset	LVGB	kNN$_S$	PAW	DACC	L++.NSE	**SAM**
SEA Concepts	**11.69**	13.83	13.39	15.68	14.48	12.50
Rotating Hyperplane	**12.53**	16.00	16.16	18.20	15.58	13.31
Moving RBF	44.84	20.36	24.04	54.34	44.50	**15.30**
Interchanging RBF	6.11	45.92	8.56	**1.40**	27.52	5.70
Moving Squares	12.17	68.87	61.01	**1.17**	65.90	2.30
Transient Chessb.	17.95	7.36	14.44	43.21	**1.98**	6.25
Mixed Drift	26.29	31.00	26.75	61.06	40.37	**13.33**
Artificial ∅	18.80	29.05	23.48	27.87	30.05	**9.81**
Artificial ∅ Rank	2.86	4.29	3.57	4.57	4.00	**1.71**
Weather	21.89	**21.53**	23.11	26.78	22.88	21.74
Electricity	**16.78**	28.61	26.13	16.87	27.24	17.52
Cover Type	9.07	**4.21**	6.76	10.05	15.00	4.8
Poker Hand	**13.65**	17.08	27.94	20.97	22.14	18.45
Outdoor	39.97	13.98	16.30	35.65	57.80	**11.25**
Rialto	39.64	22.74	24.96	28.93	40.36	**18.58**
Real world ∅	23.50	18.03	20.87	23.21	30.90	**15.40**
Real word ∅ Rank	3.17	2.33	4.00	4.17	5.33	**2.00**
Overall ∅	20.97	23.96	22.27	25.72	30.44	**12.39**
Overall ∅ Rank	3.00	3.38	3.77	4.38	4.62	**1.85**

Reasons for robustness

- Adaptation guided through error minimization
  - Dynamic size of the STM
  - Model selection for prediction
  - Reduction of hyperparameters
- Consistency between STM and LTM
- LTM acts as safety net

Q & A


Doubts
yes
Confused
Maybe
Accept
Unclear
No
Perplexed
Unsure
Reject