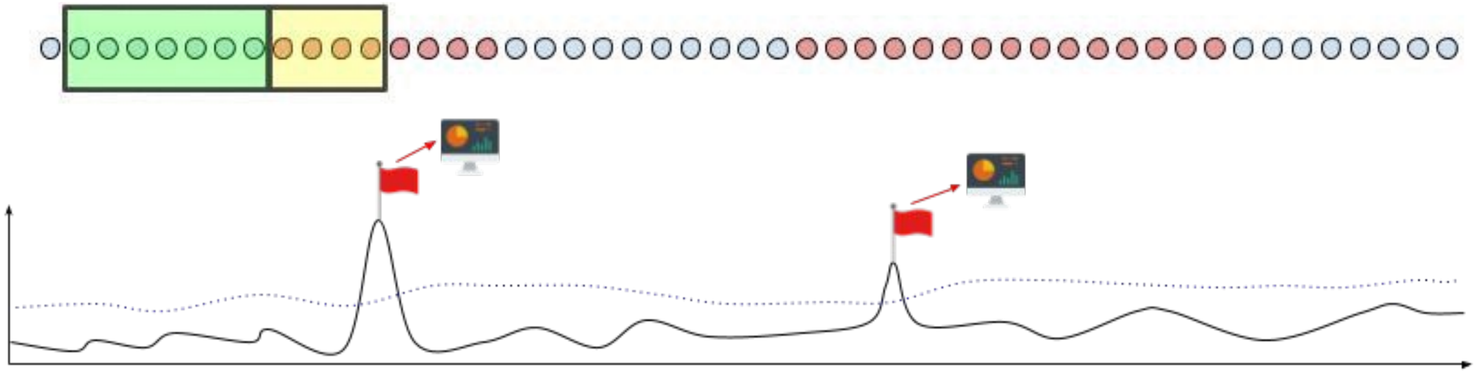
Automatic Model Monitoring for Data Streams

Pedro Bizarro
Fábio Pinto
Marco O.P. Sampaio

Based on arXiv:1908.04240 and

DSPT Meetup, 26 November 2019



Motivation

Fraud prevention

Payments processing is done in many ways ⇒ **wide scope for fraud activities**

- ATM, payment terminals
- Online
- Virtual and physical cards
- ...

feedzai

Fraud prevention as binary classification

User
name
ID
email
address
...

Temporal
timestamp
local time
expiration date
issuing date
...

Location
country
city
IP
device
...

Purchase
amount
merchant category
product
transaction type
...

Features

Fraud prevention as binary classification

Model

User	Temporal	Location	Purchase
name	timestamp	country	amount
ID	local time	city	merchant category
email	expiration date	IP	product
address	issuing date	device	transaction type
...	...	...	...

Features

Fraud prevention as binary classification

User
name
ID
email
address
...

Temporal
timestamp
local time
expiration date
issuing date
...

Location
country
city
IP
device
...

Purchase
amount
merchant category
product
transaction type
...

Features

Model →

Score
0.2

Decision
✓

Fraud prevention as binary classification

feedzai

User
name
ID
email
address
...

Temporal
timestamp
local time
expiration date
issuing date
...

Location
country
city
IP
device
...

Purchase
amount
merchant category
product
transaction type
...

Features

Model →

Score

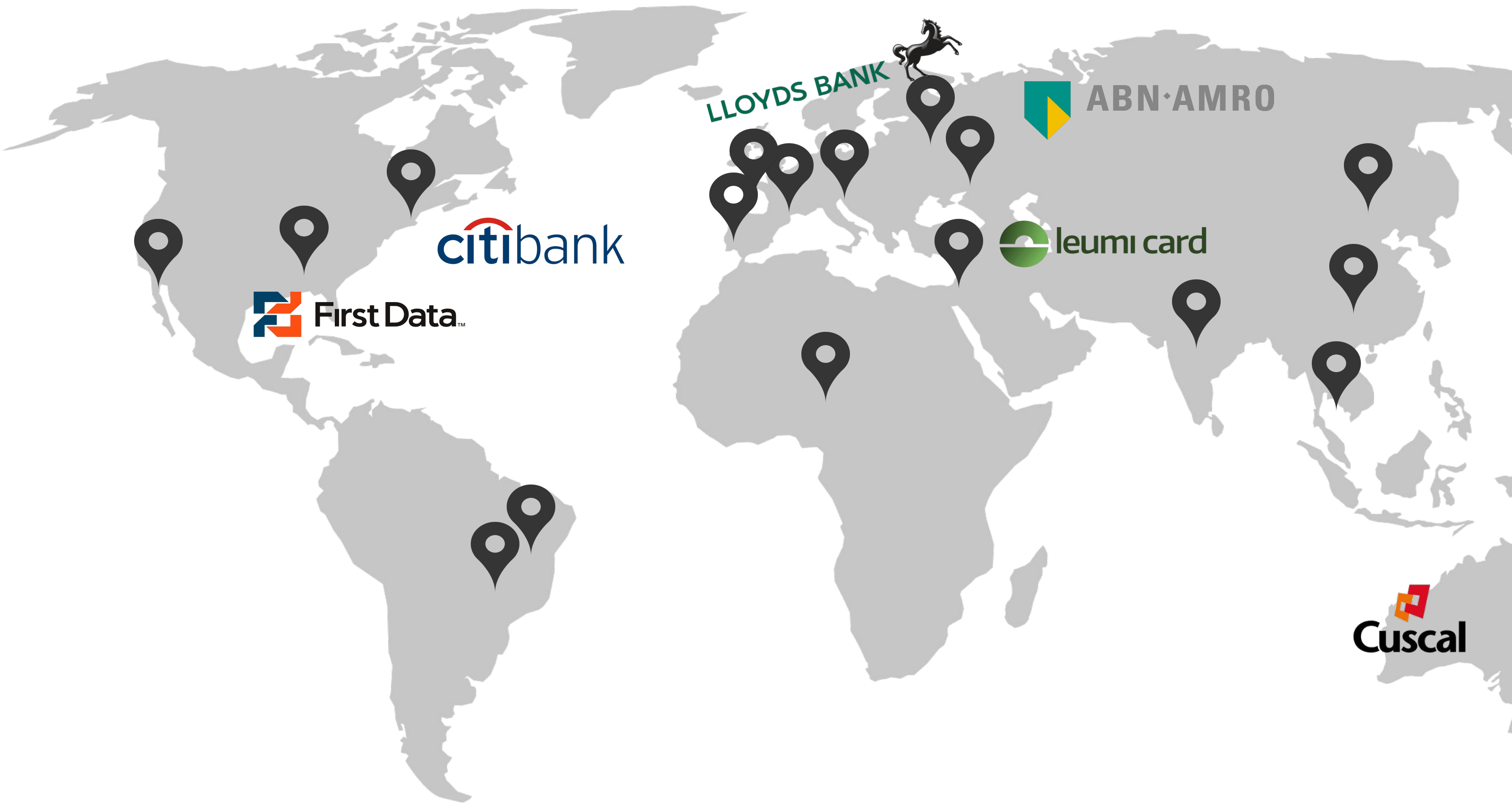
0.9

Decision

The Fraud Landscape

Examples

- Banks
- Merchants
- Payment processors



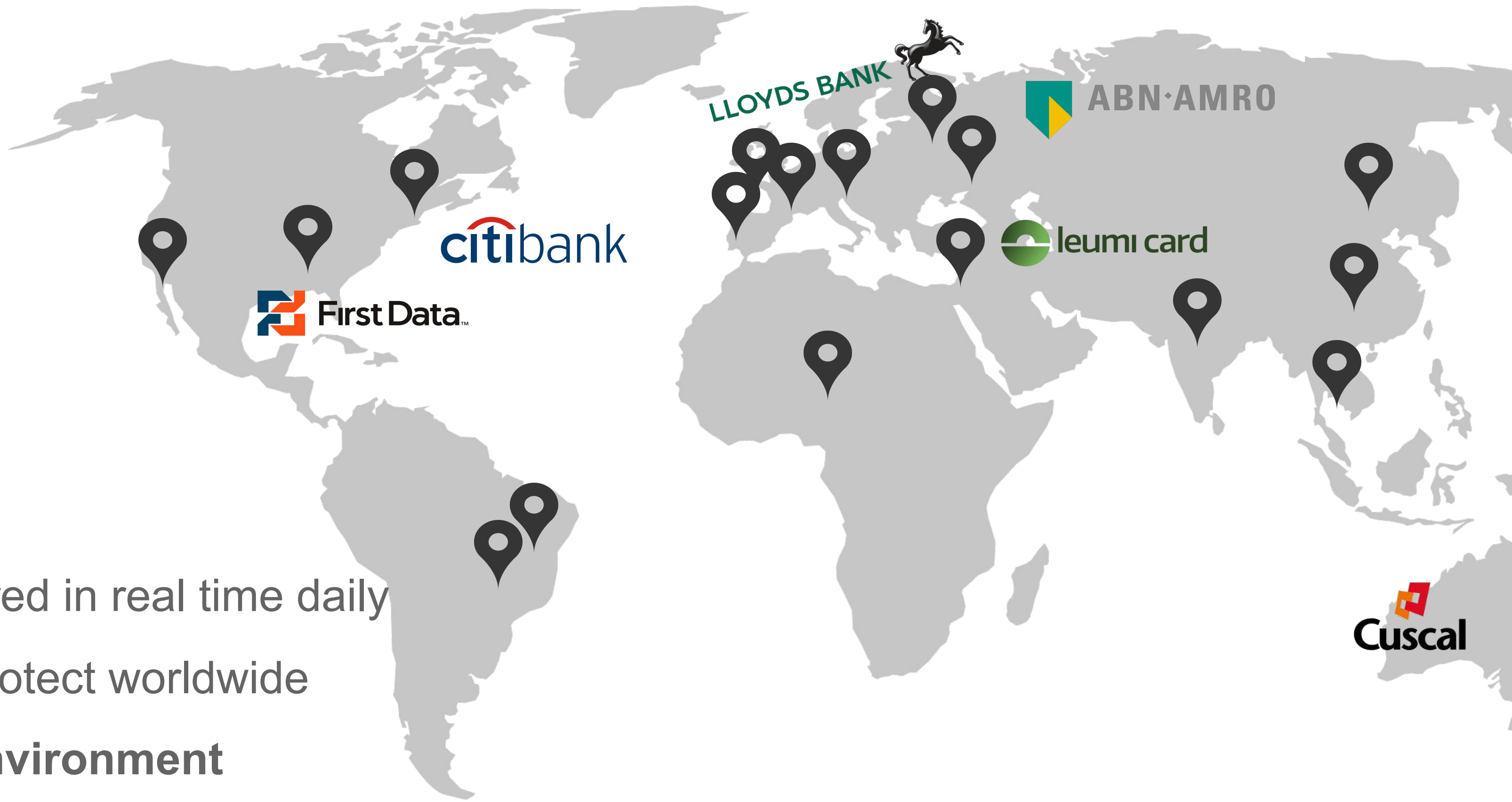
feedzai

The Fraud Landscape

Examples

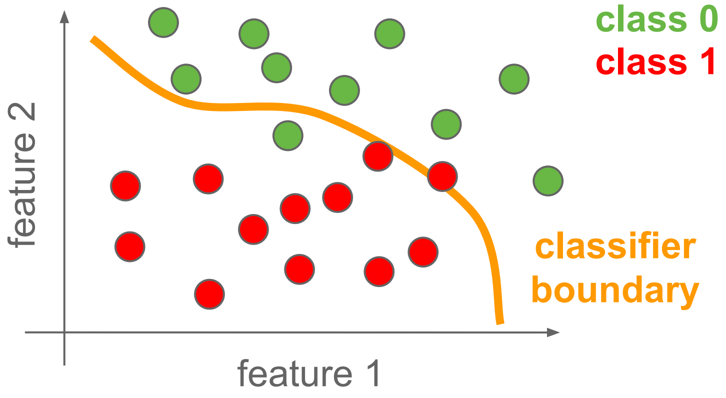
- Banks
- Merchants
- Payment processors

Characteristics



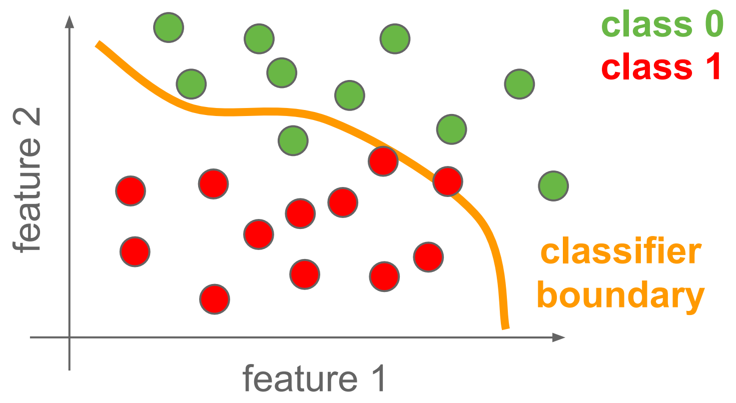
- Tens of millions of transactions scored in real time daily
- Hundreds of millions of people to protect worldwide
- **Non-stationary data streaming environment**
- **Fraudsters keep changing strategies** (adversarial)
- **Label collection can take from days, to weeks, to even months.**

Concept Drift



Concept Drift

Concept drift is a **change in the joint distribution of the data**, p(X,Y)

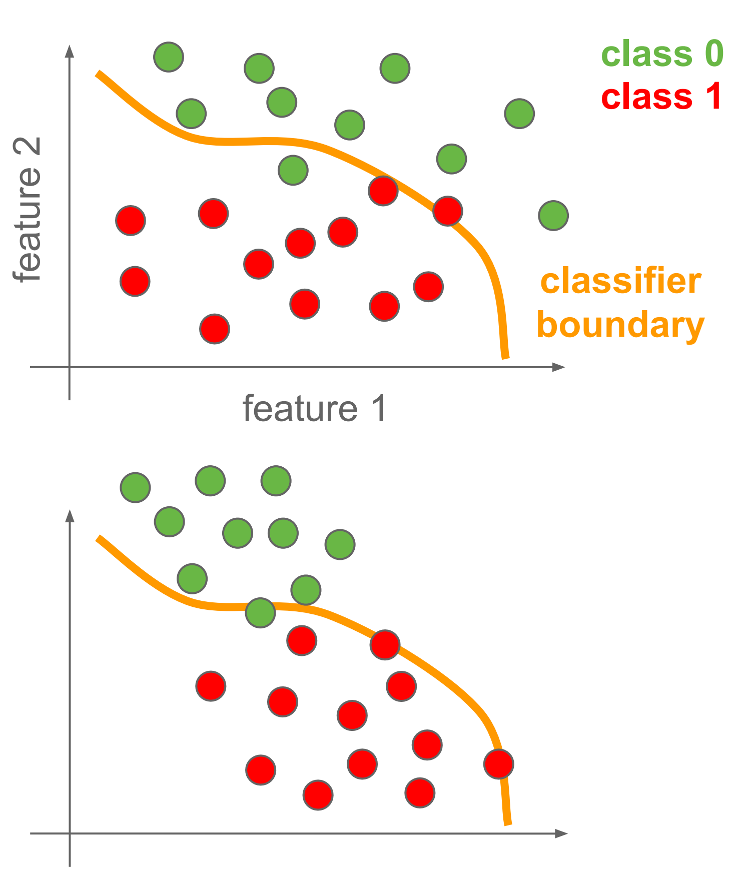


Concept Drift

Concept drift is a **change in the joint distribution of the data**, p(X,Y)

- **Virtual drift:** Distribution of features p(X) changes

  ⟶ No change in p(Y|X), i.e. relation between target Y and features X

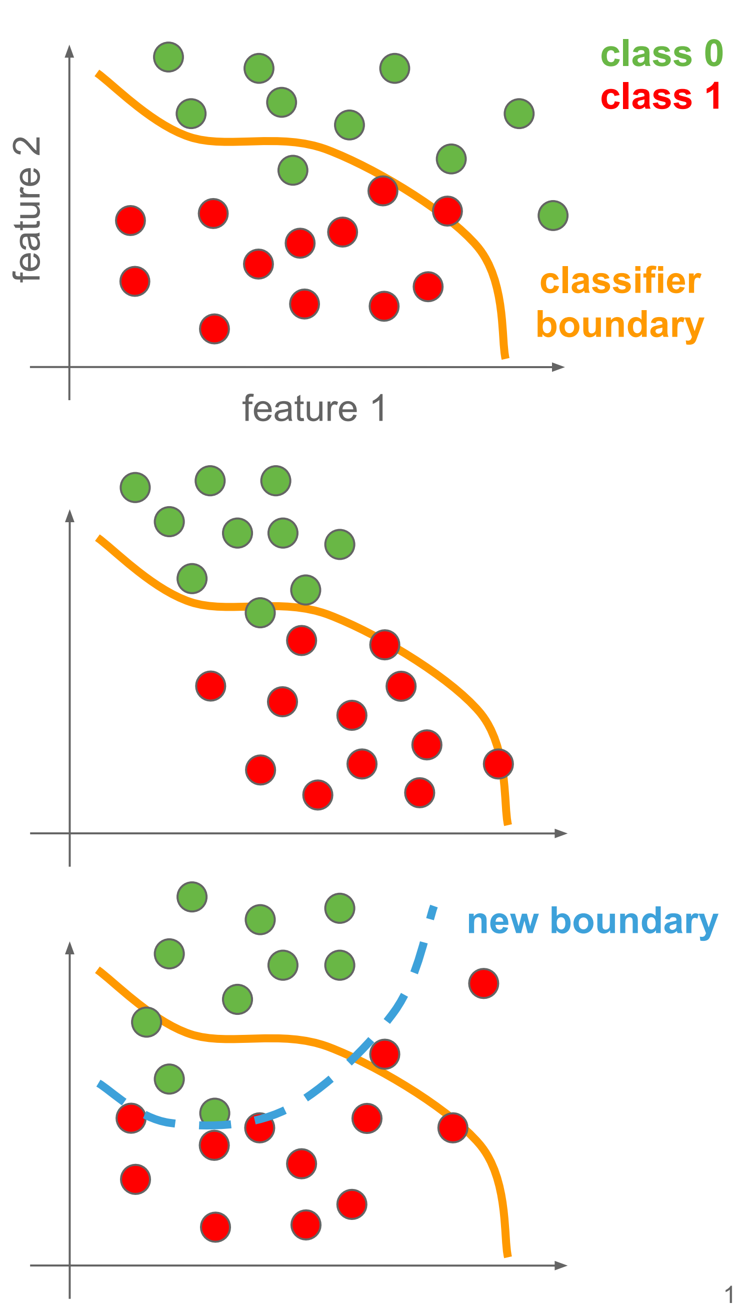


feedzai

Concept Drift

Concept drift is a **change in the joint distribution of the data**, p(X,Y)

- **Virtual drift:** Distribution of features p(X) changes
  ⟶ No change in p(Y|X), i.e. relation between target Y and features X
- **Real drift:** Only p(Y|X) changes
  ⟶ Classifier decision changes



feedzai

Concept Drift

Concept drift is a **change in the joint distribution of the data**, p(X,Y)

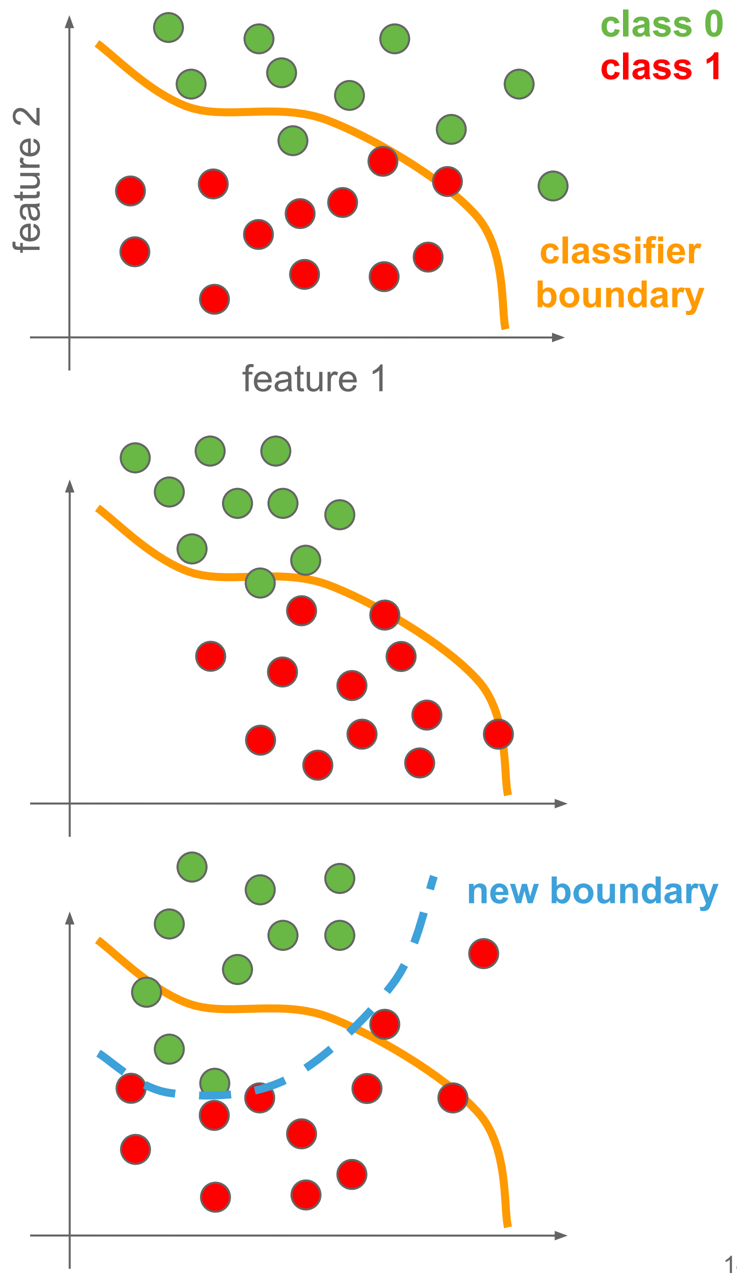
- **Virtual drift:** Distribution of features p(X) changes

  ⟶ No change in p(Y|X), i.e. relation between target Y and features X

- **Real drift:** Only p(Y|X) changes

  ⟶ Classifier decision changes

⇒ Real drift affects model performance!



Drift Detection With Delayed Labels

Most approaches in literature are **supervised**

Drift Detection With Delayed Labels

Most approaches in literature are **supervised**

But:

- Fraud labels can take weeks to arrive

- We are mostly interested in detecting sudden drifts (~ hours to days)
  E.g., new attack strategy, API changes and corrupts features.
  → In practice, long term drift in model performance is easier to detect and deal with (re-training)

feedzai

Drift Detection With Delayed Labels

Most approaches in literature are **supervised**

But:

- Fraud labels can take weeks to arrive
- We are mostly interested in detecting sudden drifts (~ hours to days)
  E.g., new attack strategy, API changes and corrupts features.
  → In practice, long term drift in model performance is easier to detect and deal with (re-training)

⇒ Can we automatically monitor the model in real time without labels?

1. Solution Overview

Example: A Bot Attack as Concept Drift

Time	Customer	Email	Card	Amount USD	Score
1:19pm	John Dow	my_dow1@mail.com	A	200.00	0.72
1:20pm	John Doe	my_dow2@mail.com	A	201.60	0.75
1:20pm	Jonny Dow	my_dow3@mail.com	A	200.00	0.76
1:20pm	J. Doe	my_dow4@mail.com	A	201.00	0.73
1:20pm	J. Dow	my_dow5@mail.com	A	200.00	0.80

Example: A Bot Attack as Concept Drift

- **Many transactions** in short time period
- Similar name, e-mail, amount, and **same card.**
- **Higher risk** scores.

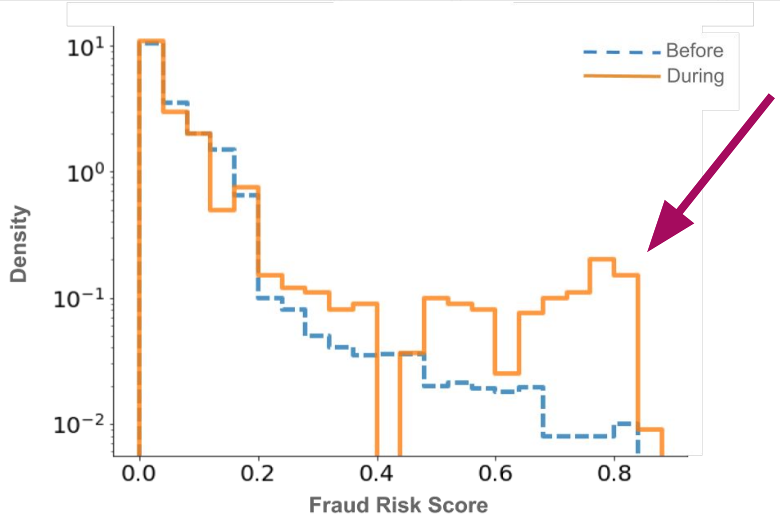
Time	Customer	Email	Card	Amount USD	Score
1:19pm	John Dow	my_dow1@mail.com	A	200.00	0.72
1:20pm	John Doe	my_dow2@mail.com	A	201.60	0.75
1:20pm	Jonny Dow	my_dow3@mail.com	A	200.00	0.76
1:20pm	J. Doe	my_dow4@mail.com	A	201.00	0.73
1:20pm	J. Dow	my_dow5@mail.com	A	200.00	0.80

Example: A Bot Attack as Concept Drift

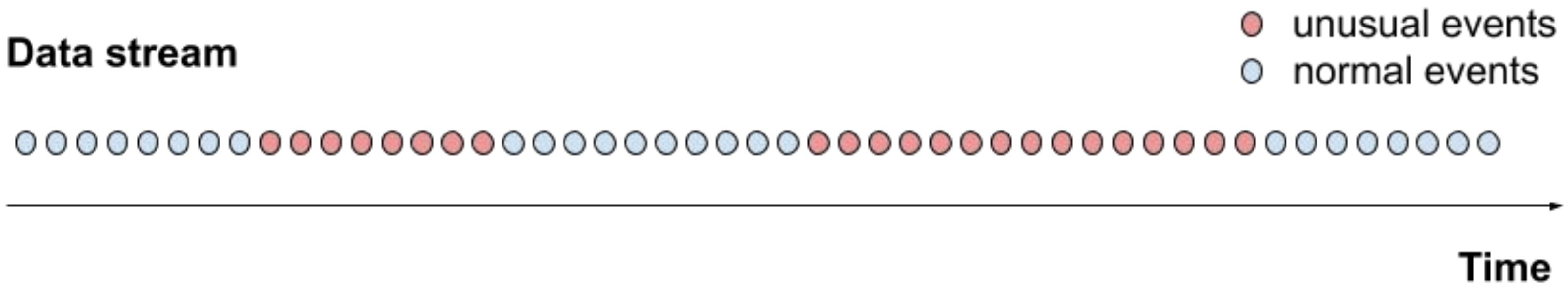
- **Many transactions** in short time period
- Similar name, e-mail, amount, and **same card.**
- **Higher risk** scores.

Time	Customer	Email	Card	Amount USD	Score
1:19pm	John Dow	my_dow1@mail.com	A	200.00	0.72
1:20pm	John Doe	my_dow2@mail.com	A	201.60	0.75
1:20pm	Jonny Dow	my_dow3@mail.com	A	200.00	0.76
1:20pm	J. Doe	my_dow4@mail.com	A	201.00	0.73
1:20pm	J. Dow	my_dow5@mail.com	A	200.00	0.80

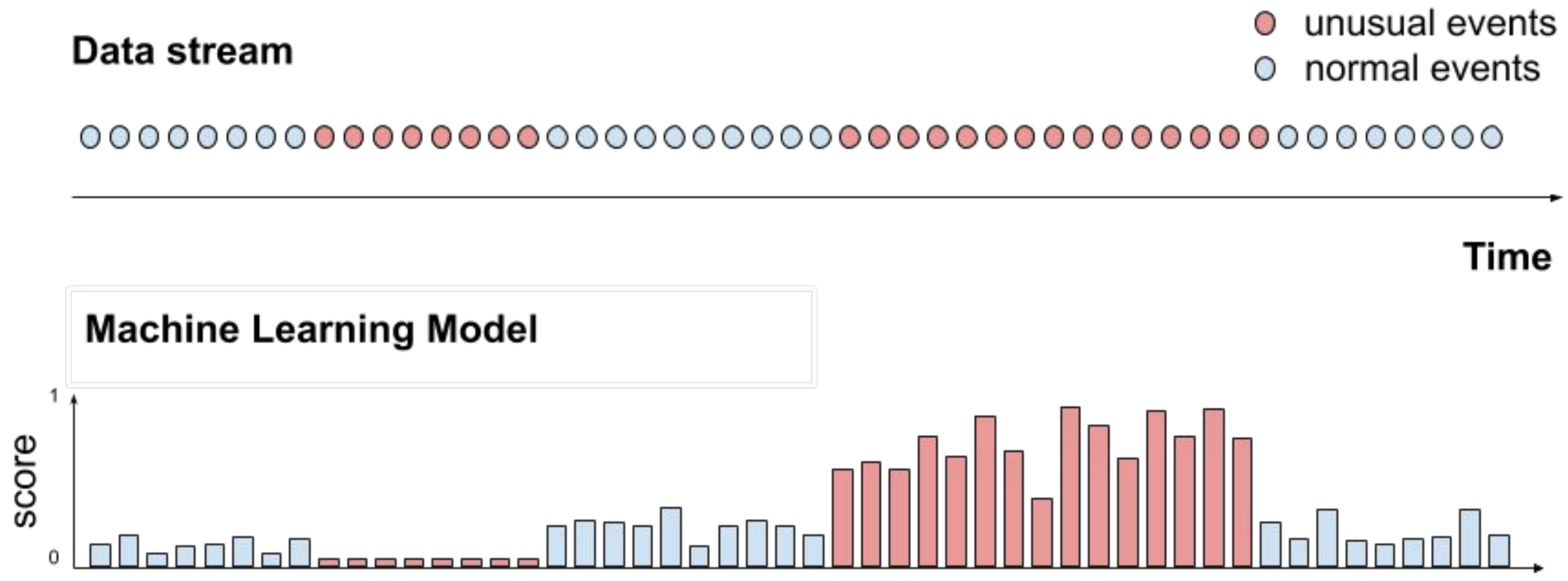
⇒ The distribution of risk scores changes.



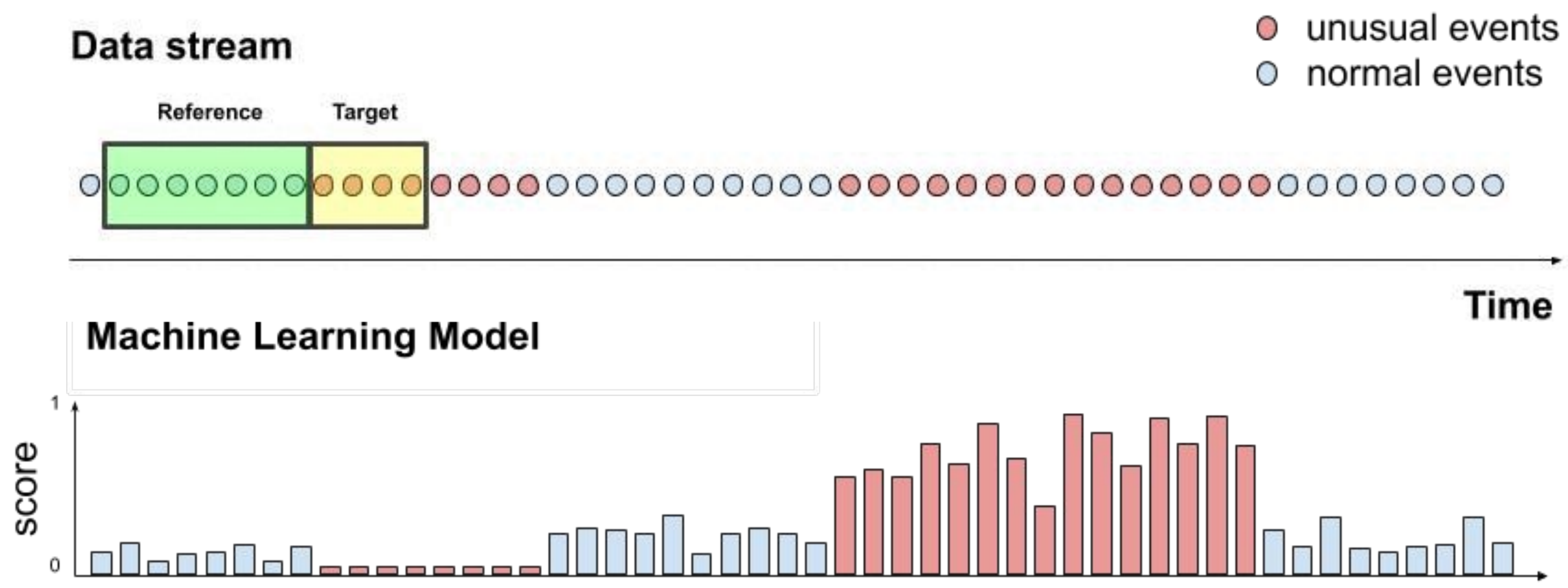
Our Solution



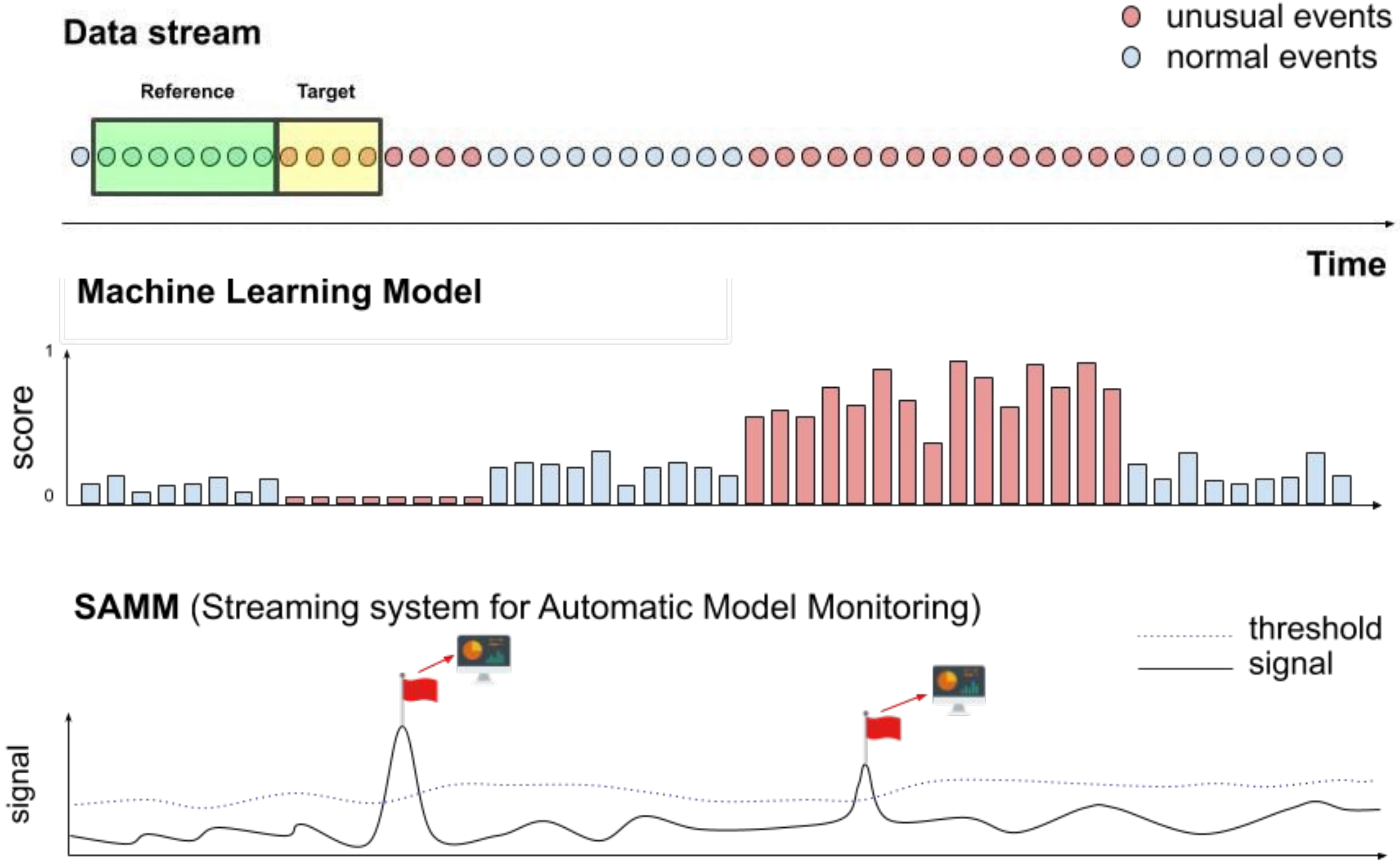
Our Solution



Our Solution



Our Solution



Summary of Requirements

SAMM (**S**treaming system for **A**utomatic **M**odel **M**onitoring)

Monitors short term drift in an unsupervised way

Provides a threshold that allows to keep false alarms under control

Provides **automatic alarm reports** with an explanation

To help a data scientist or analyst in figuring out what happened

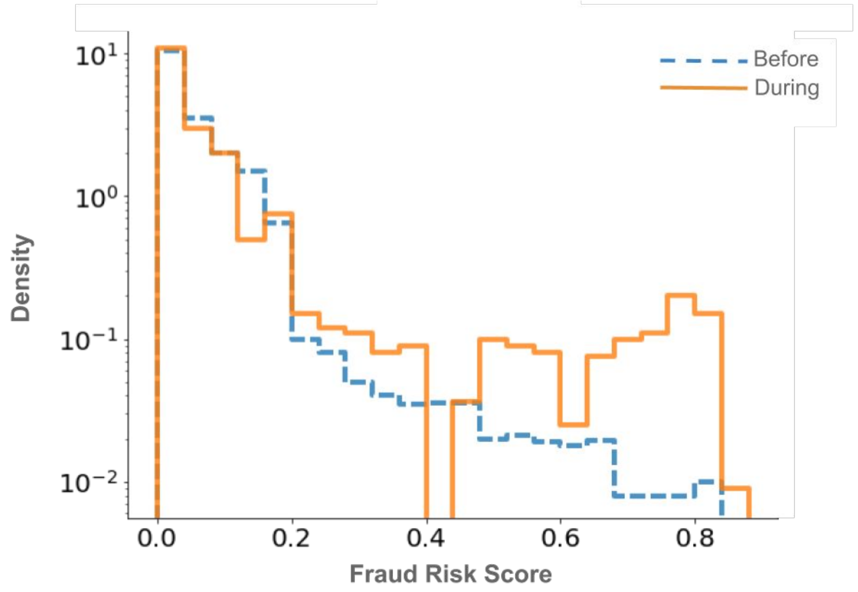
2. A Closer Look

feedzai

2a. The Signal

Measure of dissimilarity between distributions:

- Jensen-Shannon Divergence (JSD)
- Kolmogorov-Smirnov
- Anderson-Darling
- Kuiper's

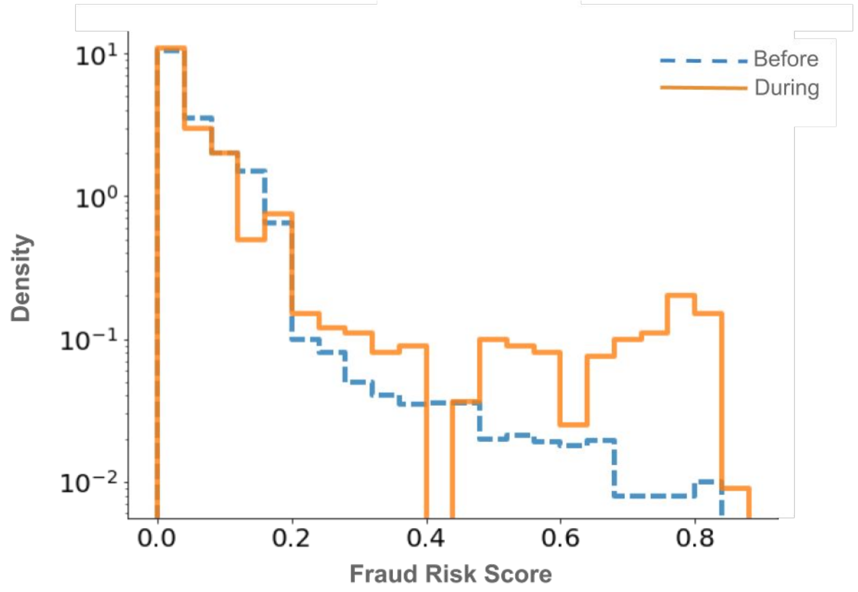


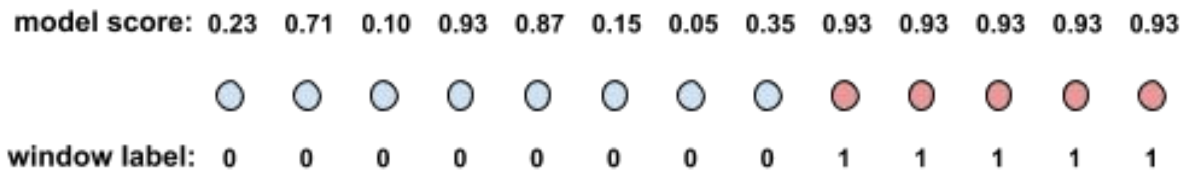
feedzai

2a. The Signal

Measure of dissimilarity between distributions:

- Jensen-Shannon Divergence (JSD)
- Kolmogorov-Smirnov
- Anderson-Darling
- Kuiper's

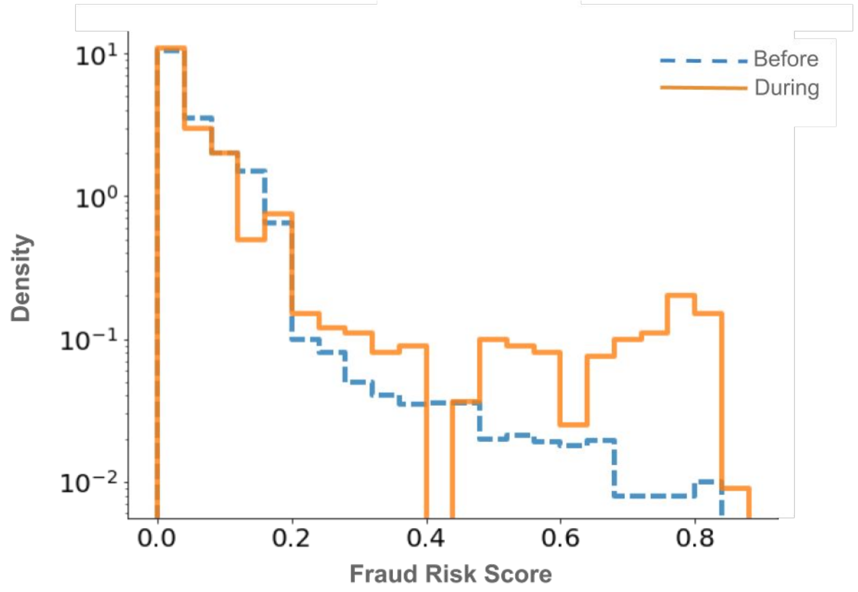


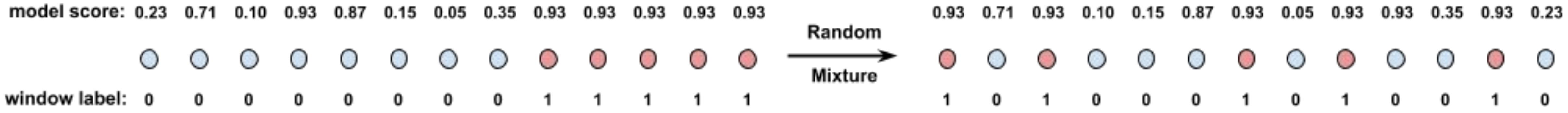


2a. The Signal

Measure of dissimilarity between distributions:

- Jensen-Shannon Divergence (JSD)
- Kolmogorov-Smirnov
- Anderson-Darling
- Kuiper's

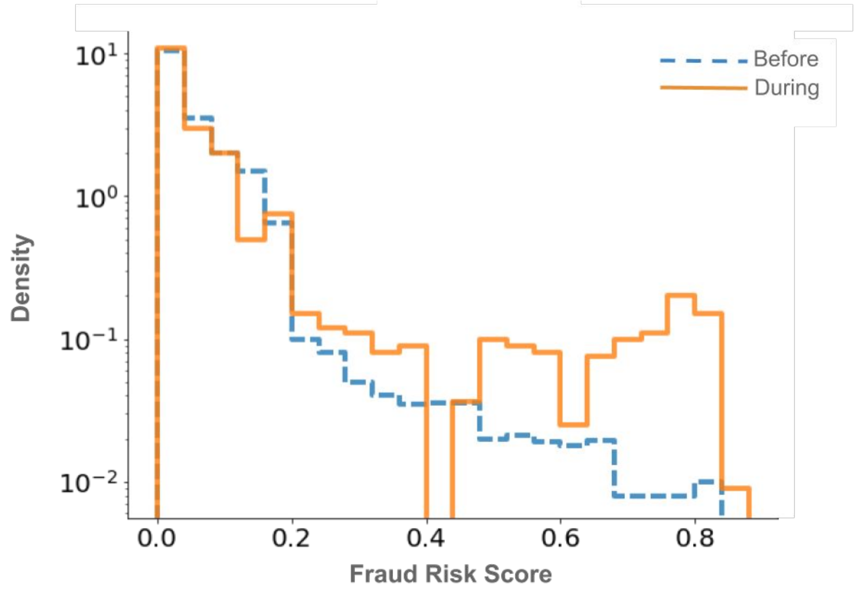


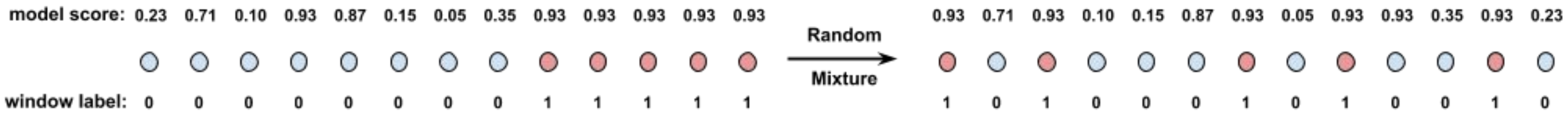


2a. The Signal

Measure of dissimilarity between distributions:

- Jensen-Shannon Divergence (JSD)
- Kolmogorov-Smirnov
- Anderson-Darling
- Kuiper's





JSD = **Mutual information between** **model score** and **window label**

2b. The Threshold

Main ingredients

SPEAR (**S**treaming **P**ercentiles **E**stim**A**to**R** of past signal values)

- Stochastic approximation with single pass over the data
- Constant memory

feedzai

2b. The Threshold

Main ingredients

SPEAR (**S**treaming **P**ercentiles **E**stim**A**to**R** of past signal values)

- Stochastic approximation with single pass over the data
- Constant memory

AdaSPEAR (forgets older signal values)

2b. The Threshold

Main ingredients

SPEAR (**S**treaming **P**ercentiles **E**stim**A**to**R** of past signal values)

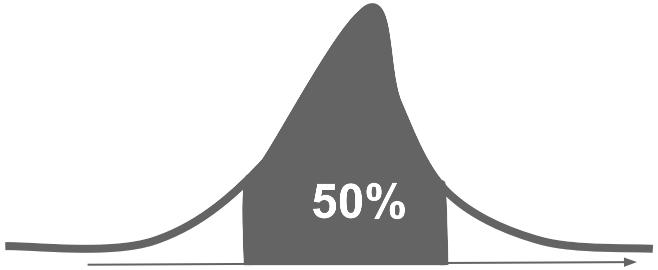
- Stochastic approximation with single pass over the data
- Constant memory

AdaSPEAR (forgets older signal values)

Tukey's definition of **outlier**

Q1 and Q3 are percentiles 25 and 75

T = **Q3** + K (**Q3** - **Q1**)



2b. The Threshold

Main ingredients

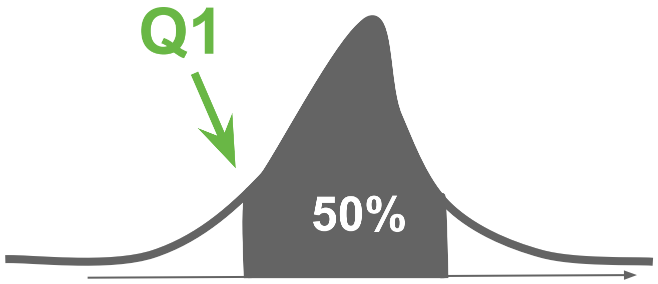
SPEAR (**S**treaming **P**ercentiles **E**stim**A**to**R** of past signal values)

- Stochastic approximation with single pass over the data
- Constant memory

AdaSPEAR (forgets older signal values)

Tukey's definition of **outlier**

Q1 and Q3 are percentiles 25 and 75



$$T = Q3 + K\,(Q3 - Q1)$$

2b. The Threshold

Main ingredients

SPEAR (**S**treaming **P**ercentiles **E**stim**A**to**R** of past signal values)

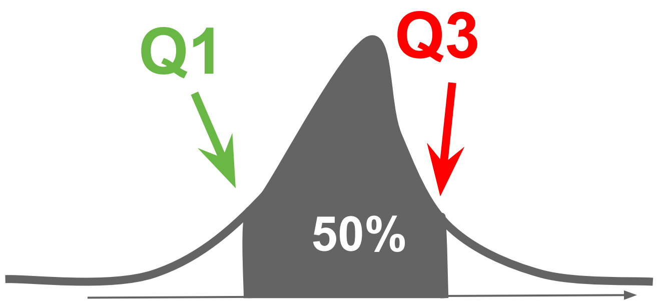
- Stochastic approximation with single pass over the data
- Constant memory

AdaSPEAR (forgets older signal values)

Tukey’s definition of **outlier**

Q1 and Q3 are percentiles 25 and 75

$$T = Q3 + K\,(Q3 - Q1)$$



feedzai

2b. The Threshold

Main ingredients

SPEAR (**S**treaming **P**ercentiles **E**stim**A**to**R** of past signal values)

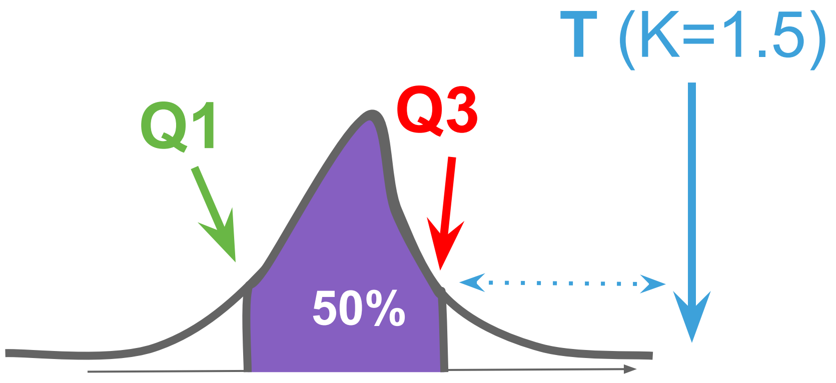
- Stochastic approximation with single pass over the data
- Constant memory

AdaSPEAR (forgets older signal values)

Tukey's definition of **outlier**

Q1 and Q3 are percentiles 25 and 75

$$T = Q3 + K\,(Q3 - Q1)$$



feedzai

2b. The Threshold

Main ingredients

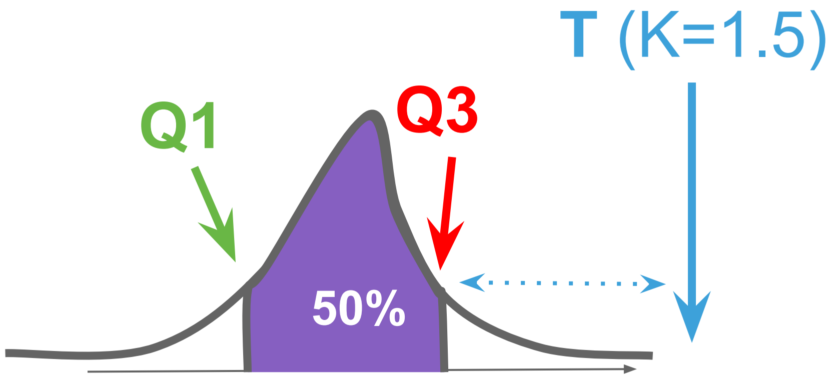
SPEAR (**S**treaming **P**ercentiles **E**stim**A**to**R** of past signal values)

- Stochastic approximation with single pass over the data
- Constant memory

AdaSPEAR (forgets older signal values)

Tukey’s definition of **outlier**

Q1 and Q3 are percentiles 25 and 75

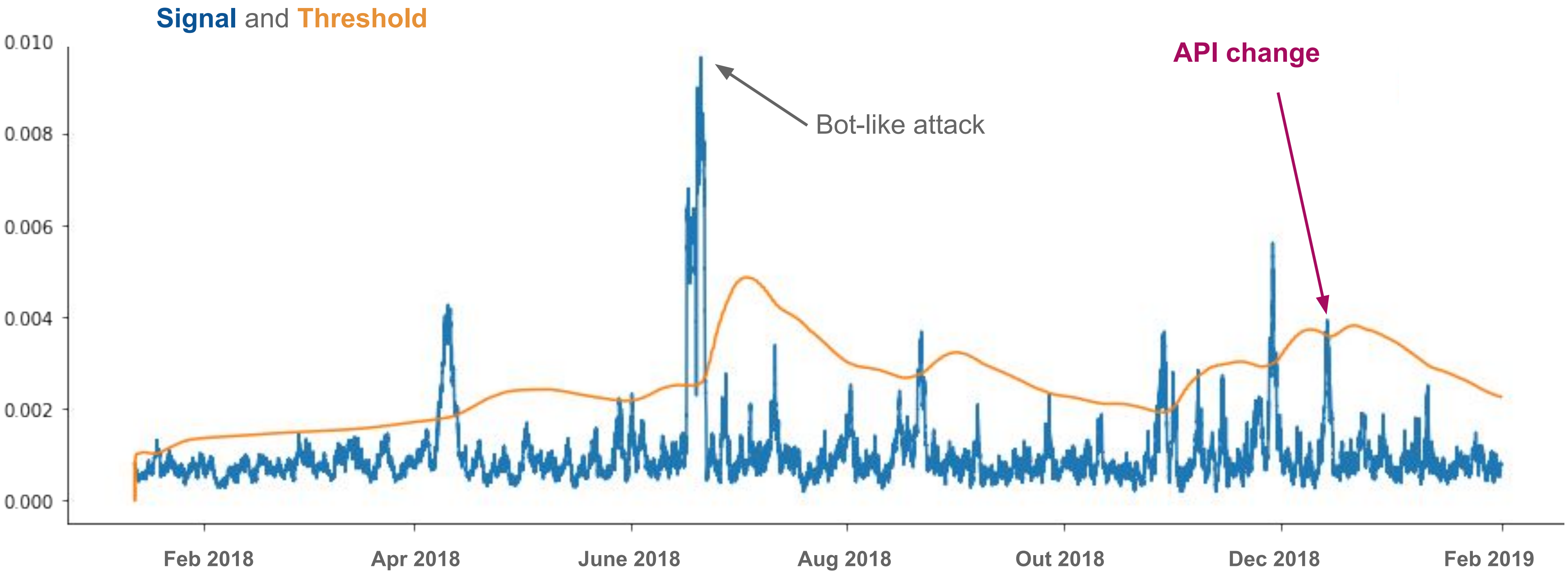


$$T = Q3 + K\,(Q3 - Q1)$$

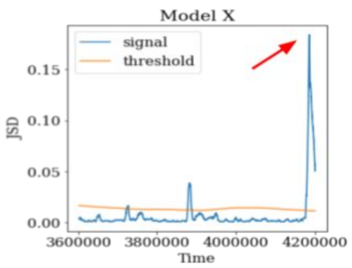
\+ **Delayed** smoothing

2b. The Threshold

Delayed Adaptive Threshold



Alarm Trigger Details



Reference window start:	2/03/2017 12:34
Reference window end:	5/03/2017 08:23
Target window start:	5/03/2017 08:23
Target window end:	5/03/2017 18:23

Alarm Report

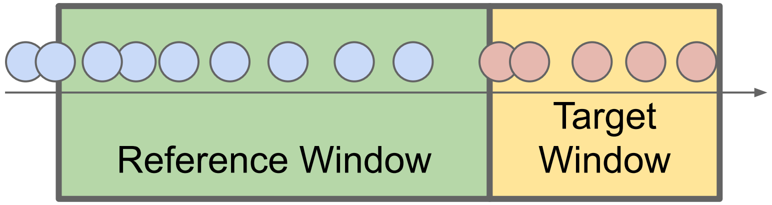
Top ranked events responsible for the alarm

feedzai

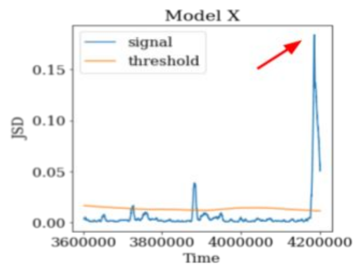
2c. Alarm report

For each alarm:

- Label **reference = 0**, and **target = 1**
- **Train GBDT model** to learn pattern that splits transactions:
  a. Use **drift score to rank transactions** in target window
  b. Use **feature importance to rank features**



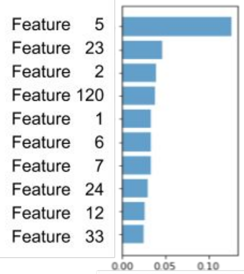
Alarm Trigger Details



Reference window start: 2/03/2017 12:34
Reference window end: 5/03/2017 08:23

Target window start: 5/03/2017 08:23
Target window end: 5/03/2017 18:23

Alarm Report



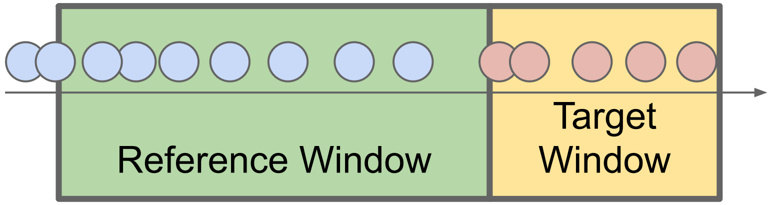
Feature importance (top 10)

Top ranked events responsible for the alarm										
transaction id	F5	F23	F2	F120	F1	F6	F7	F24	F12	F33
234	0.13	1.24	0	-23.00	4.45	-0.29	1	1	0.00	0
3432	0.14	1.24	1	0.14	1.24	0.14	0	0	0.00	1
212	9.24	3.56	1	9.24	3.56	9.24	0	0	1.33	1
867	9.24	3.56	0	0.14	0.14	0.14	0	0	0.00	0
436	3.56	217.83	0	0.23	-3242	0.23	1	1	-1.20	0
964	999.00	0.14	1	-23.00	4.45	56345	1	1	0.00	1
748	-32.42	0.23	0	0.23	0.14	1.24	1	1	0.00	0

2c. Alarm report

For each alarm:

- Label **reference = 0**, and **target = 1**
- **Train GBDT model** to learn pattern that splits transactions:
  a. Use **drift score to rank transactions** in target window
  b. Use **feature importance to rank features**

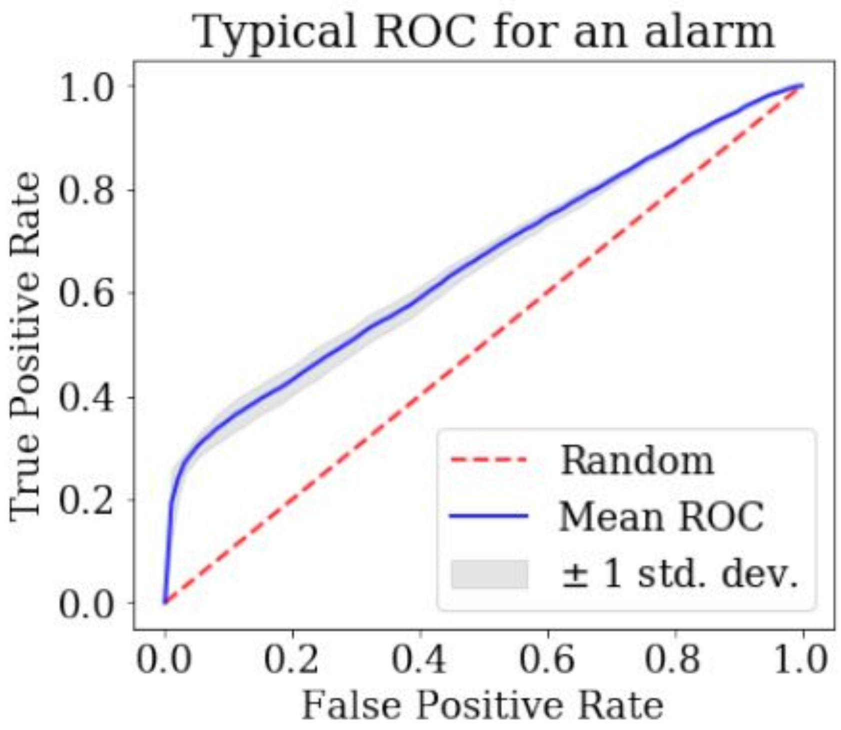


But:

- **Danger** of achieving **perfect split due to time correlated features**!
- **Solution:** Eliminate time-correlated features.

2c. Alarm report

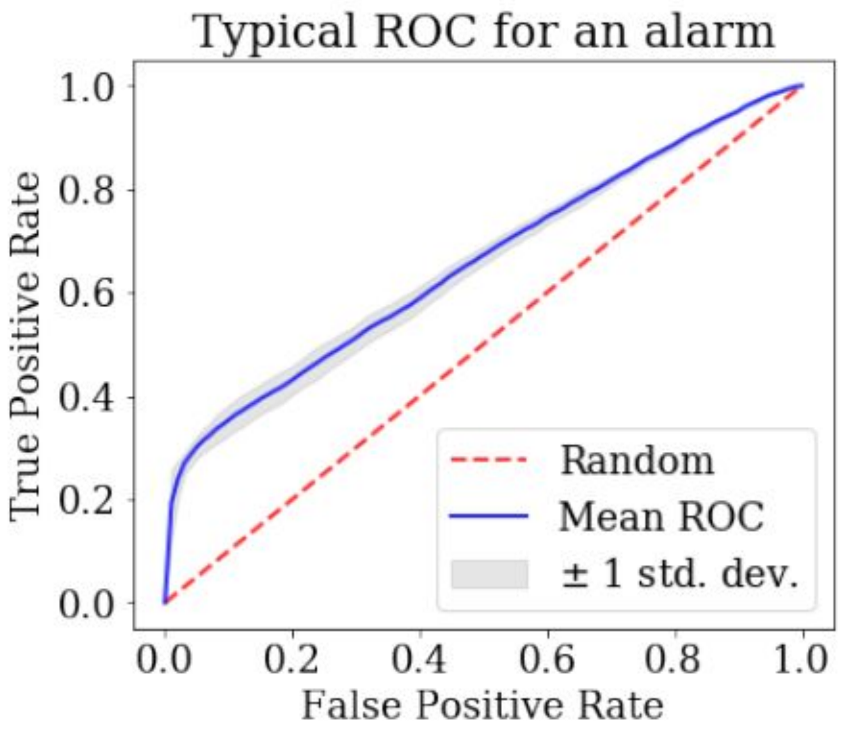
Cross Validation

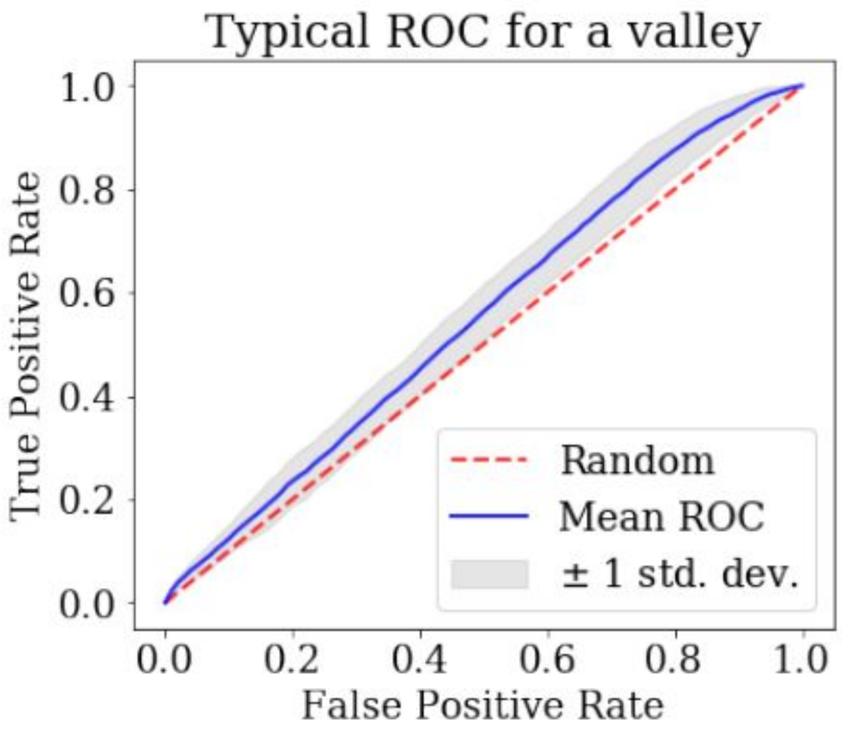


For true alarms the drift model finds a pattern

2c. Alarm report

Cross Validation

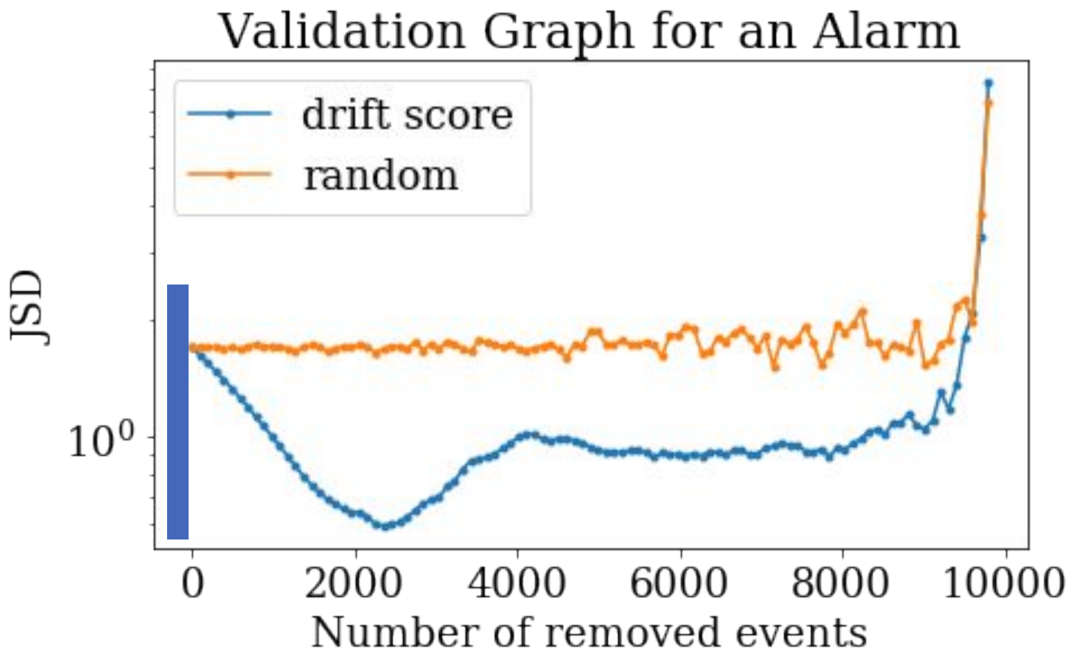




For true alarms the drift model finds a pattern whereas for valleys it does not.

2c. Alarm report

Ranking Validation



Top ranked events responsible for the alarm

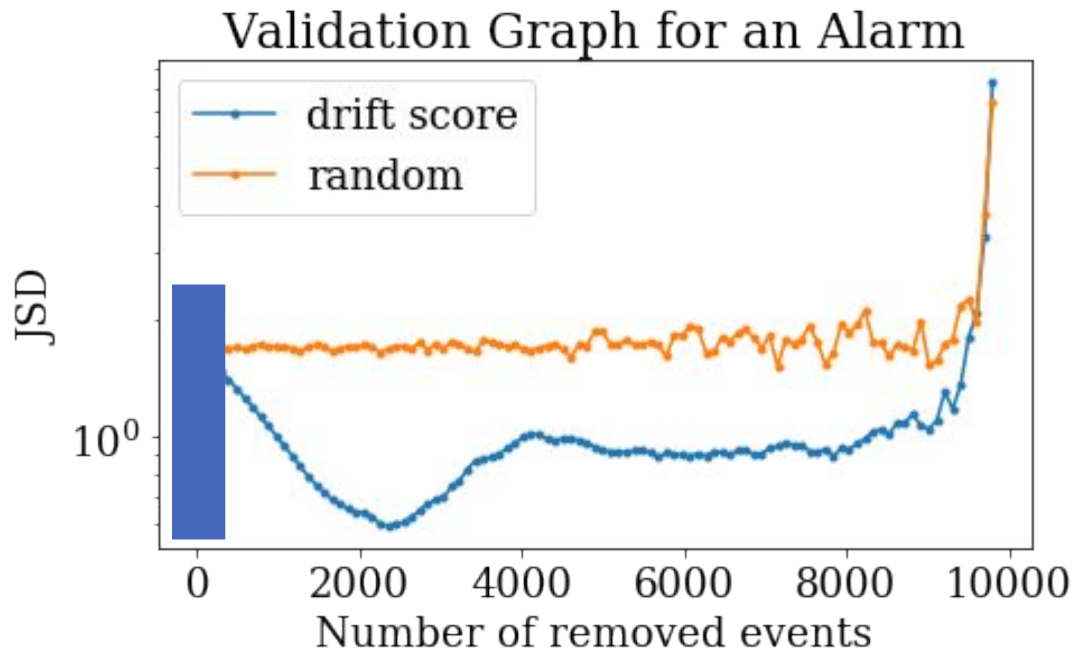
transaction id	F5	F23	F2	F120	F1	F6	F7	F24	F12	F33
234	0.13	1.24	0	-23.00	4.45	-0.29	1	1	0.00	0
3432	0.14	1.24	1	0.14	1.24	0.14	0	0	0.00	1
212	9.24	3.56	1	9.24	3.56	9.24	0	0	1.33	1
867	9.24	3.56	0	0.14	0.14	0.14	0	0	0.00	0
436	3.56	217.83	0	0.23	-3242	0.23	1	1	-1.20	0
964	999.00	0.14	1	-23.00	4.45	56345	1	1	0.00	1
748	-32.42	0.23	0	0.23	0.14	1.24	1	1	0.00	0

(...)

234	0.13	1.24	0	-23.00	4.45	-0.29	1	1	0.00	0
3432	0.14	1.24	1	0.14	1.24	0.14	0	0	0.00	1
212	9.24	3.56	1	9.24	3.56	9.24	0	0	1.33	1
867	9.24	3.56	0	0.14	0.14	0.14	0	0	0.00	0
436	3.56	217.83	0	0.23	-3242	0.23	1	1	-1.20	0
964	999.00	0.14	1	-23.00	4.45	56345	1	1	0.00	1
748	-32.42	0.23	0	0.23	0.14	1.24	1	1	0.00	0

2c. Alarm report

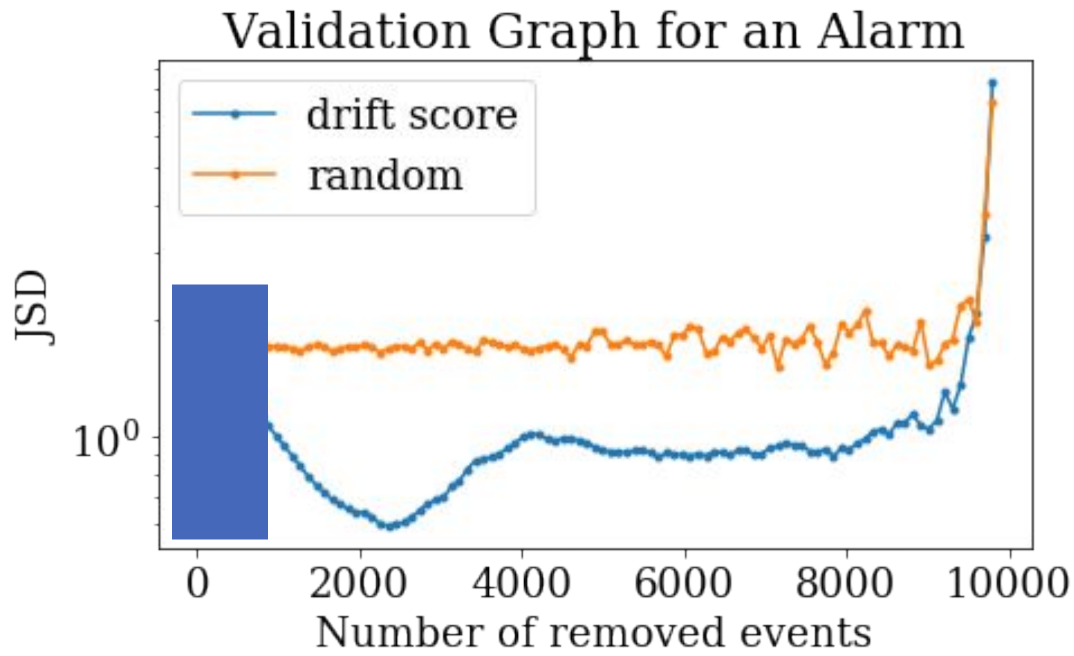
Ranking Validation



Top ranked events responsible for the alarm										
transaction id	F5	F23	F2	F120	F1	F6	F7	F24	F12	F33
867	9.24	3.56	0	0.14	0.14	0.14	0	0	0.00	0
436	3.56	217.83	0	0.23	-3242	0.23	1	1	-1.20	0
964	999.00	0.14	1	-23.00	4.45	56345	1	1	0.00	1
748	-32.42	0.23	0	0.23	0.14	1.24	1	1	0.00	0
(...)										
234	0.13	1.24	0	-23.00	4.45	-0.29	1	1	0.00	0
3432	0.14	1.24	1	0.14	1.24	0.14	0	0	0.00	1
212	9.24	3.56	1	9.24	3.56	9.24	0	0	1.33	1
867	9.24	3.56	0	0.14	0.14	0.14	0	0	0.00	0
436	3.56	217.83	0	0.23	-3242	0.23	1	1	-1.20	0
964	999.00	0.14	1	-23.00	4.45	56345	1	1	0.00	1
748	-32.42	0.23	0	0.23	0.14	1.24	1	1	0.00	0

2c. Alarm report

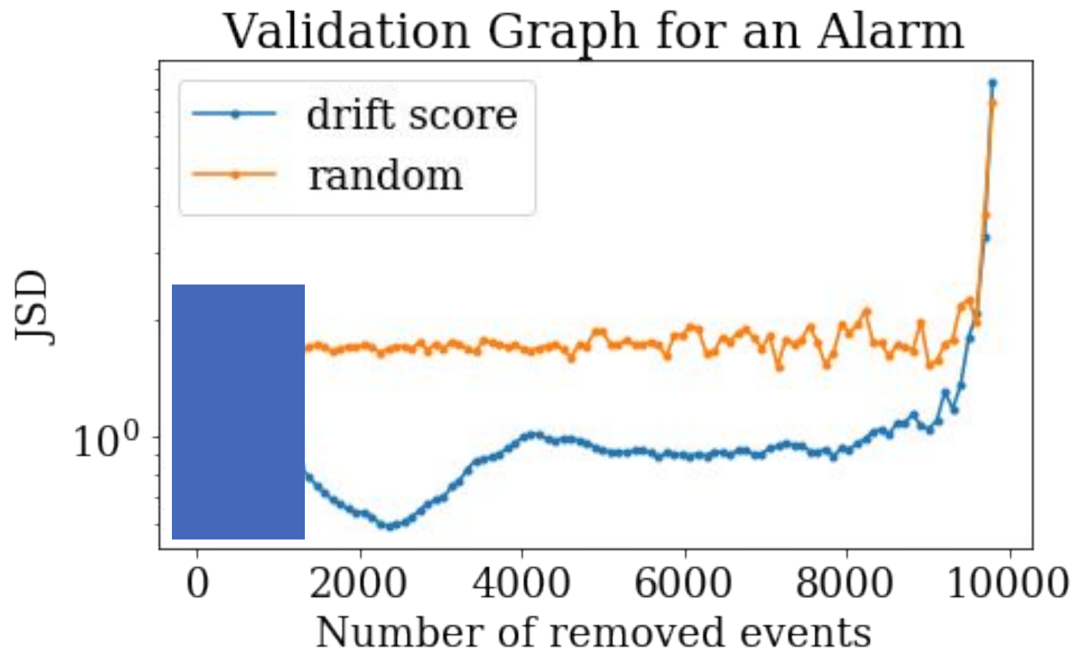
Ranking Validation



Top ranked events responsible for the alarm										
transaction id	F5	F23	F2	F120	F1	F6	F7	F24	F12	F33
748	-32.42	0.23	0	0.23	0.14	1.24	1	1	0.00	0
(...)										
234	0.13	1.24	0	-23.00	4.45	-0.29	1	1	0.00	0
3432	0.14	1.24	1	0.14	1.24	0.14	0	0	0.00	1
212	9.24	3.56	1	9.24	3.56	9.24	0	0	1.33	1
867	9.24	3.56	0	0.14	0.14	0.14	0	0	0.00	0
436	3.56	217.83	0	0.23	-3242	0.23	1	1	-1.20	0
964	999.00	0.14	1	-23.00	4.45	56345	1	1	0.00	1
748	-32.42	0.23	0	0.23	0.14	1.24	1	1	0.00	0

2c. Alarm report

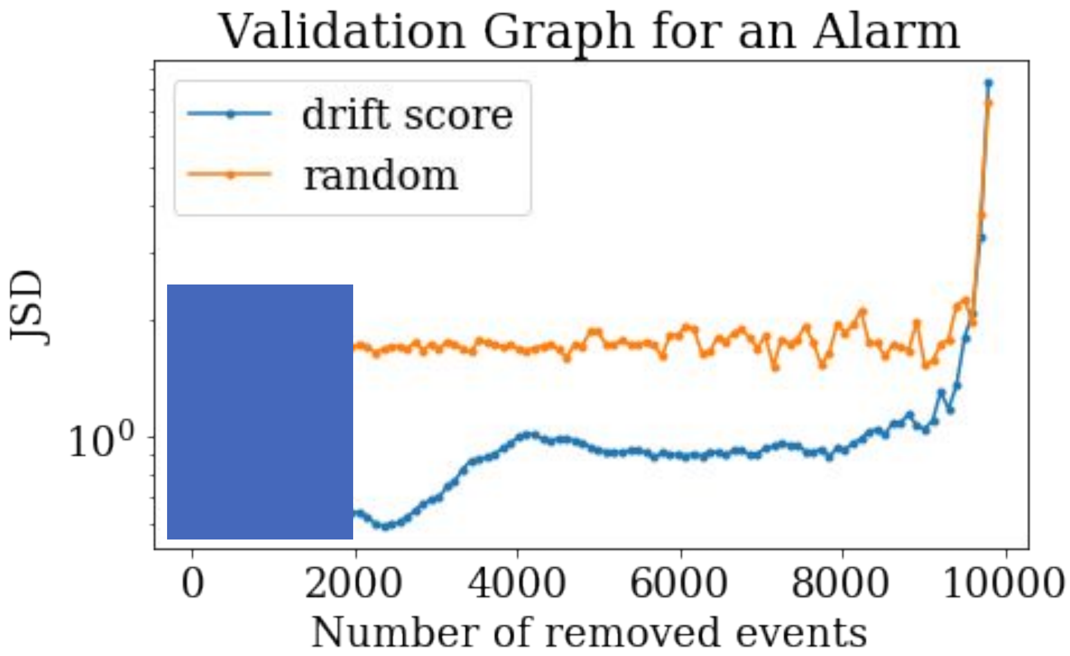
Ranking Validation



Top ranked events responsible for the alarm										
transaction id	F5	F23	F2	F120	F1	F6	F7	F24	F12	F33
3432	0.14	1.24	1	0.14	1.24	0.14	0	0	0.00	1
212	9.24	3.56	1	9.24	3.56	9.24	0	0	1.33	1
867	9.24	3.56	0	0.14	0.14	0.14	0	0	0.00	0
436	3.56	217.83	0	0.23	-3242	0.23	1	1	-1.20	0
964	999.00	0.14	1	-23.00	4.45	56345	1	1	0.00	1
748	-32.42	0.23	0	0.23	0.14	1.24	1	1	0.00	0

2c. Alarm report

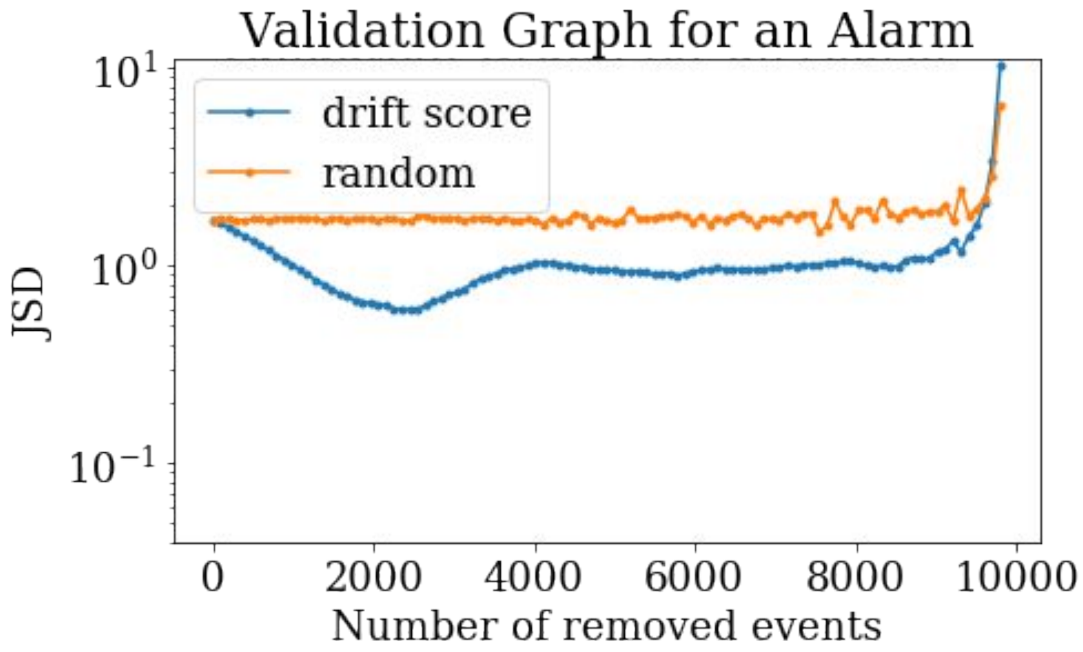
Ranking Validation



Top ranked events responsible for the alarm										
transaction id	F5	F23	F2	F120	F1	F6	F7	F24	F12	F33
[redacted]										
964	999.00	0.14	1	-23.00	4.45	56345	1	1	0.00	1
748	-32.42	0.23	0	0.23	0.14	1.24	1	1	0.00	0

2c. Alarm report

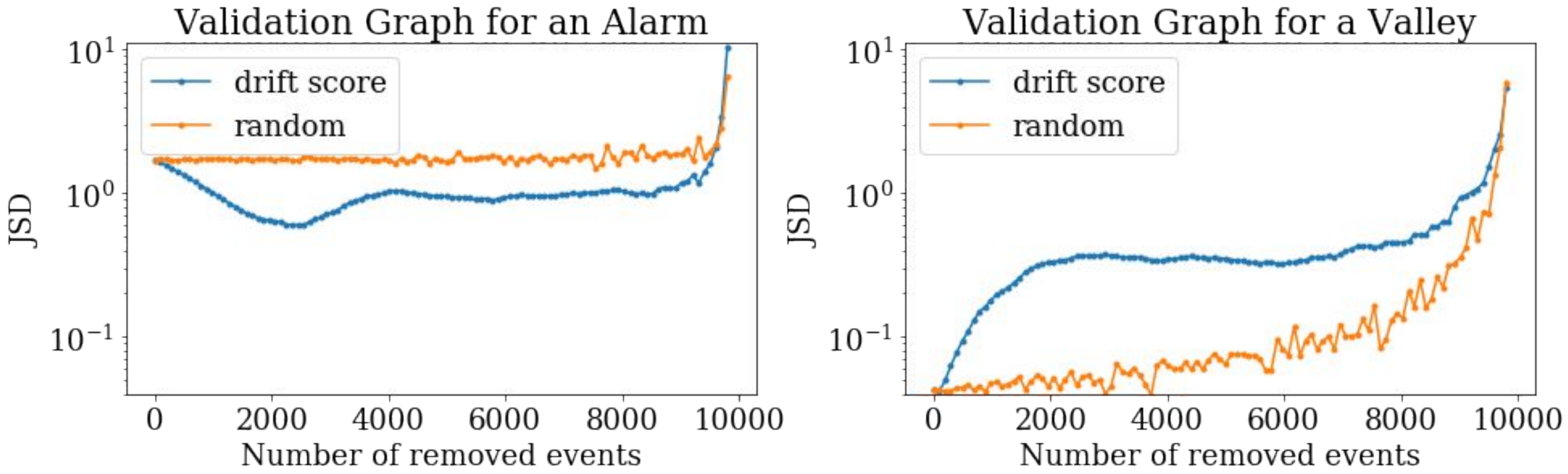
Ranking Validation



Removing top ranked transactions lowers signal for alarms

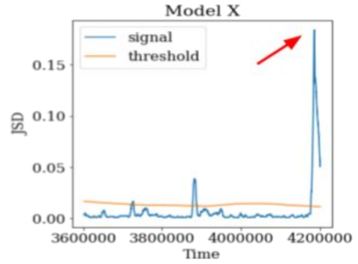
2c. Alarm report

Ranking Validation



Removing top ranked transactions lowers signal for alarms but not for valleys.

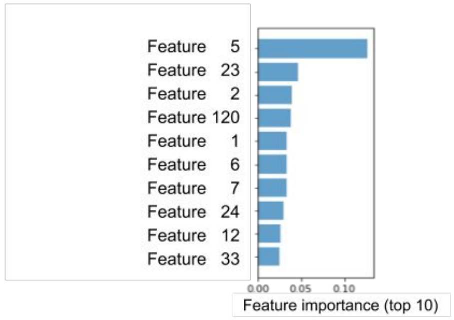
Alarm Trigger Details

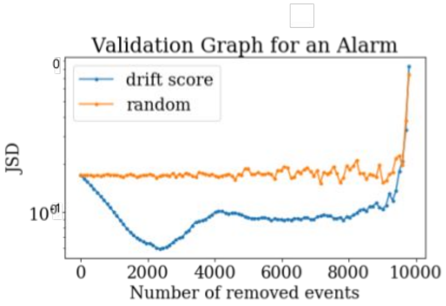


Reference window start: 2/03/2017 12:34
Reference window end: 5/03/2017 08:23

Target window start: 5/03/2017 08:23
Target window end: 5/03/2017 18:23

Alarm Report





Top ranked events responsible for the alarm

transaction id	F5	F23	F2	F120	F1	F6	F7	F24	F12	F33
234	0.13	1.24	0	-23.00	4.45	-0.29	1	1	0.00	0
3432	0.14	1.24	1	0.14	1.24	0.14	0	0	0.00	1
212	9.24	3.56	1	9.24	3.56	9.24	0	0	1.33	1
867	9.24	3.56	0	0.14	0.14	0.14	0	0	0.00	0
436	3.56	217.83	0	0.23	-3242	0.23	1	1	-1.20	0
964	999.00	0.14	1	-23.00	4.45	56345	1	1	0.00	1
748	-32.42	0.23	0	0.23	0.14	1.24	1	1	0.00	0

3. Experiments

The Datasets

Dataset	Features	Days	Transactions	Transactions per day
$A1$	213	212	1,046,482	4936
$A2$	213	212	2,667,548	12,583
$A3$	213	212	4,945,509	23,328
$B1$	279	229	4,401,807	19,221
$B2$	279	229	9,229,013	40,301

Table 1: Summary statistics for each dataset-region.

Experimental Design

- We **generated 100 reports** (20 per dataset-regions)
- We **mixed** a selection of **true alarms and valleys**
- We had two data scientists for each dataset-region (10 reports each).

Experimental Design

- We **generated 100 reports** (20 per dataset-regions)
- We **mixed** a selection of **true alarms and valleys**
- We had two data scientists for each dataset-region (10 reports each).

For each report, they were asked to **rate from 1 to 5**:

1. How **confident** are you that this is a **true alarm**?
2. How clearly can you **identify a pattern** in the transactions?
3. After looking at the **validation plot**, please provide a new answer to 1.
   (the validation plot was hidden from the user for 1 and 2).

Experimental Design

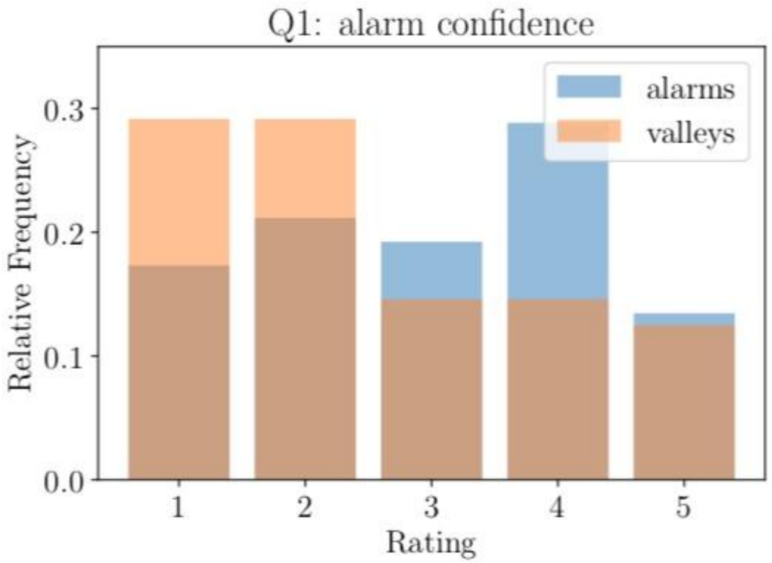
- We **generated 100 reports** (20 per dataset-regions)
- We **mixed** a selection of **true alarms and valleys**
- We had two data scientists for each dataset-region (10 reports each).

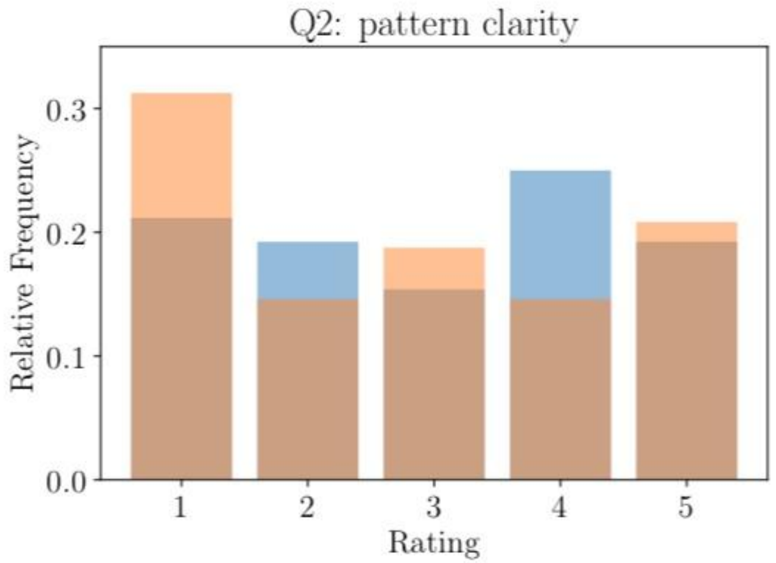
For each report, they were asked to **rate from 1 to 5**:

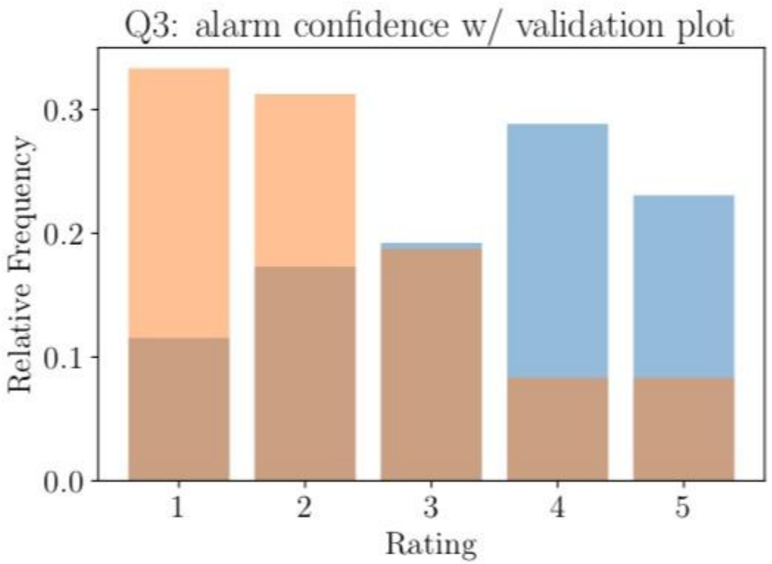
1. How **confident** are you that this is a **true alarm**?
2. How clearly can you **identify a pattern** in the transactions?
3. After looking at the **validation plot**, please provide a new answer to 1.
   (the validation plot was hidden from the user for 1 and 2).

Hypothesis: Reports based on alarms have higher ratings for all questions

Results







	Q1	Q2	Q3
Aggregated	**0.037**	0.220	**5.6e-5**
Dataset A	**0.037**	**0.037**	**0.003**
Dataset B	0.370	0.822	**0.006**

Table 2: P-values for the various Mann-Whitney U tests (with Holm-Bonferroni correction for comparisons by dataset). For all tests, the α-level was set at 0.05. Values in bold represent tests where the p-value was smaller than α.

4. Conclusions

Main Takeaways

- We proposed **SAMM,** a system to **monitor ML models for data streams**.

- It **detects drift in an unsupervised way**, computing a signal and threshold with an **efficient percentiles estimation algorithm** (SPEAR/AdaSPEAR),

- It provides an explanation **report that domain experts consider useful**.

Future directions

- Study streams with high seasonality (not an issue for datasets in our paper)
- Study effect of contents presented in the alarm reports from a UX perspective

feedzai

Future directions

- Study streams with high seasonality (not an issue for datasets in our paper)
- Study effect of contents presented in the alarm reports from a UX perspective

THANK YOU

Topics for discussion session:

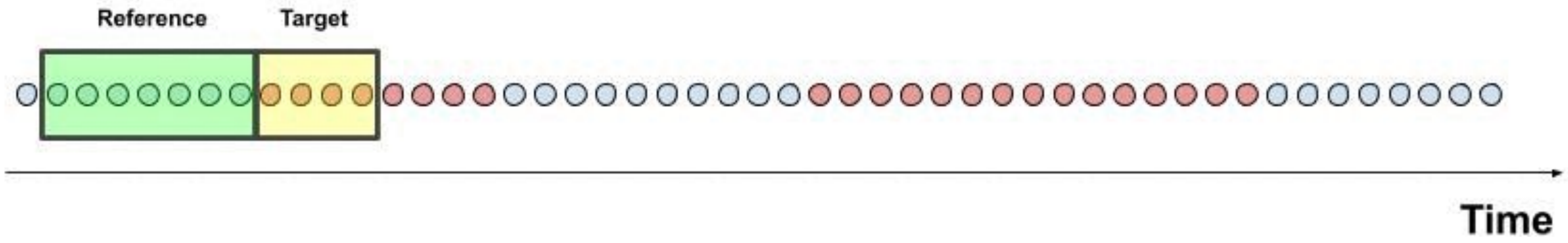
- Topics above (UX & Seasonality)
- Challenges in online concept drift detection.
- Virtual concept drift (multiple testing) and cases when it could matter.
- Supervised concept drift detection and long term drift.

Bonus Slides

2a. The Signal

Window Configurations

Contiguous Windows:

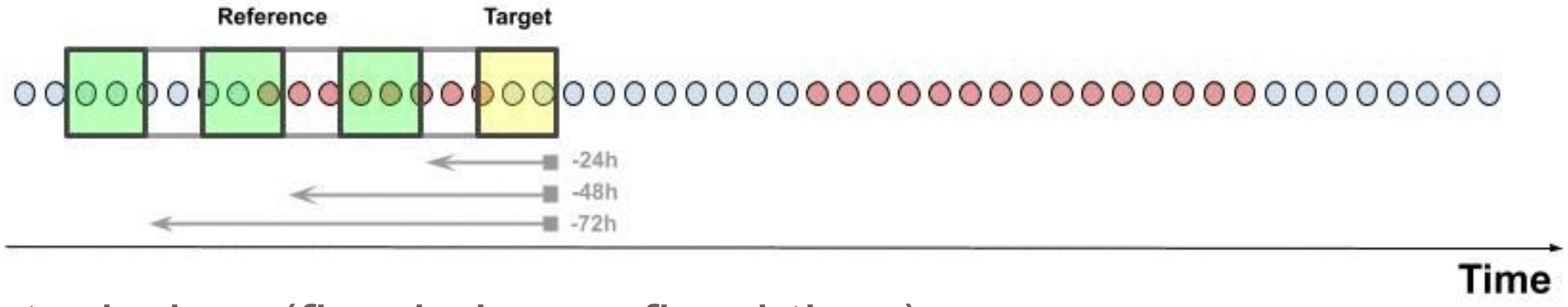


- Two consecutive windows (to detect sudden drifts)
- Preferably fixed size (to control statistics) but can be fixed time
- Size chosen to monitor some period (e.g., the last 6 hours)

2a. The Signal

Window Configurations

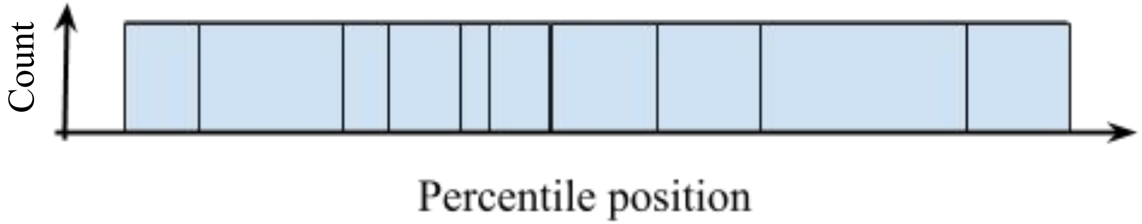
Homologous Windows:



- One target window (fixed size or fixed time)
- Reference replicas in same periods as target but, e.g., on three previous days
- Suitable to remove seasonalities in the signal

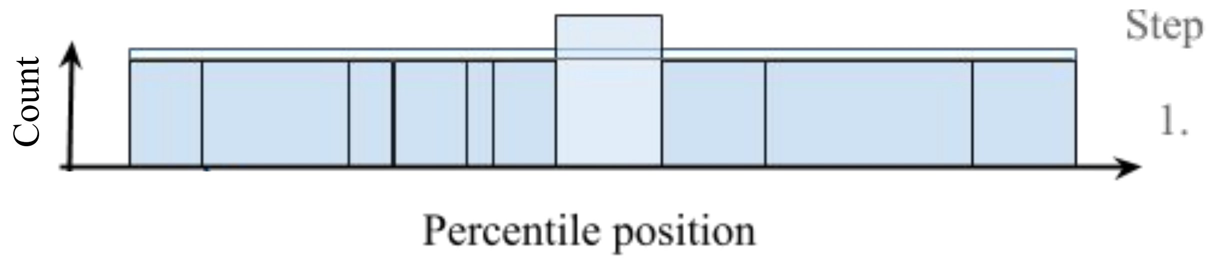
2b. The Threshold

SPEAR (Streaming Percentiles EstimAtoR)



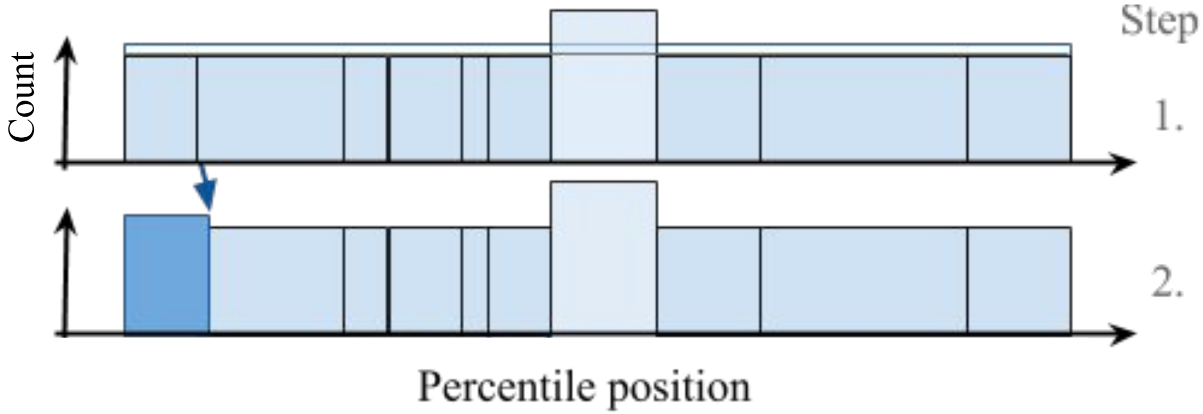
2b. The Threshold

SPEAR (Streaming Percentiles EstimAtoR)



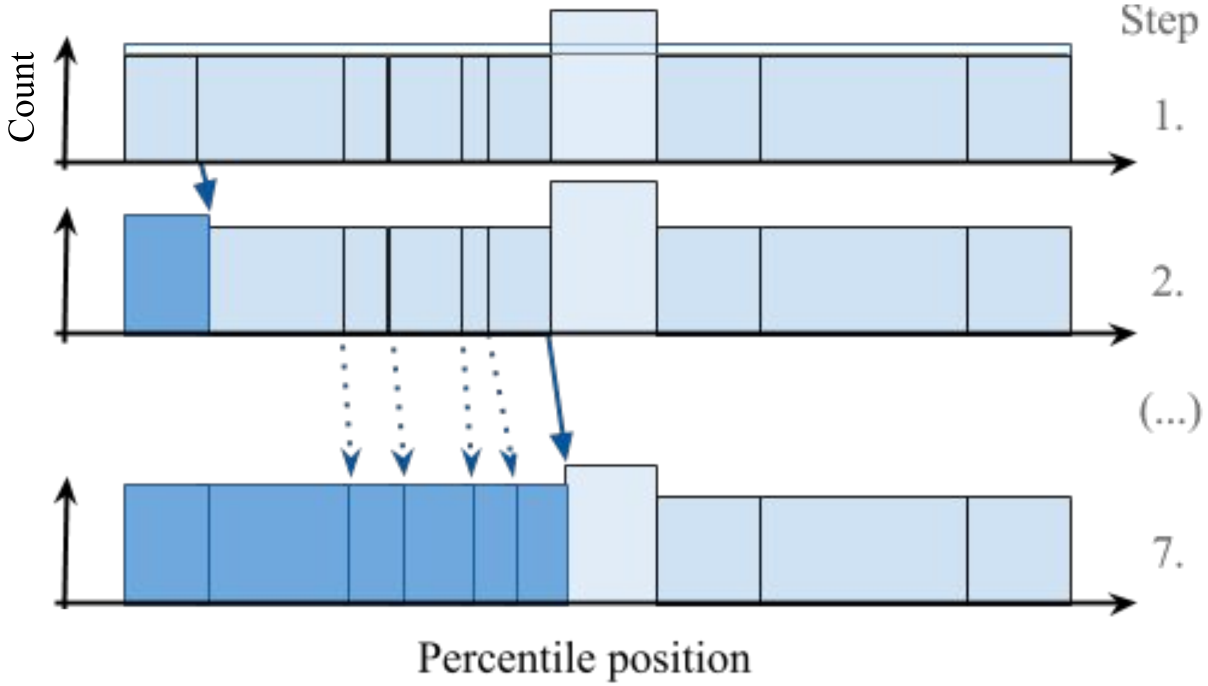
2b. The Threshold

SPEAR (Streaming Percentiles EstimAtoR)



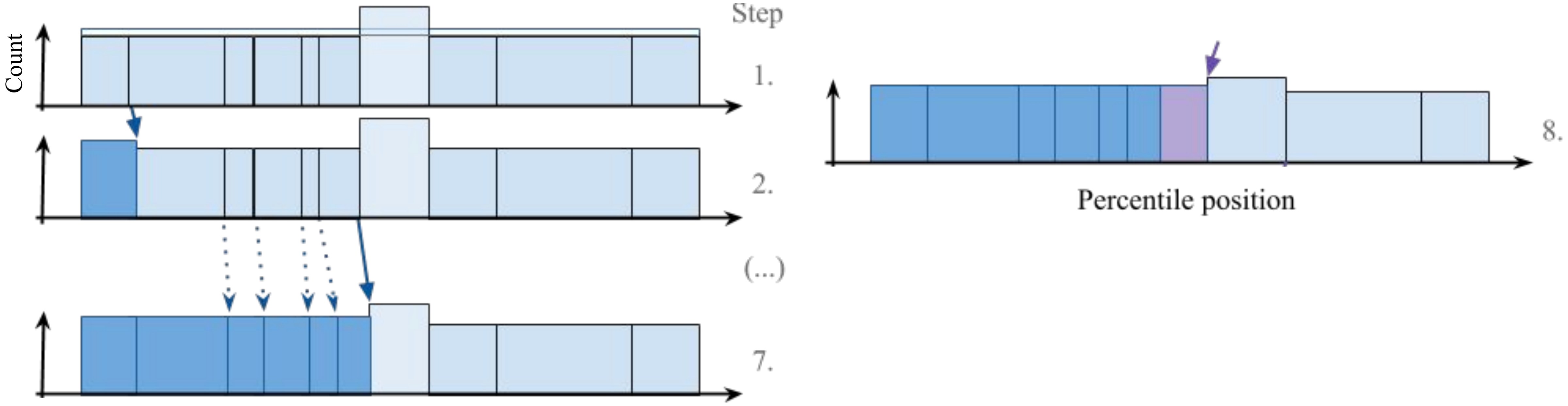
2b. The Threshold

SPEAR (Streaming Percentiles EstimAtoR)



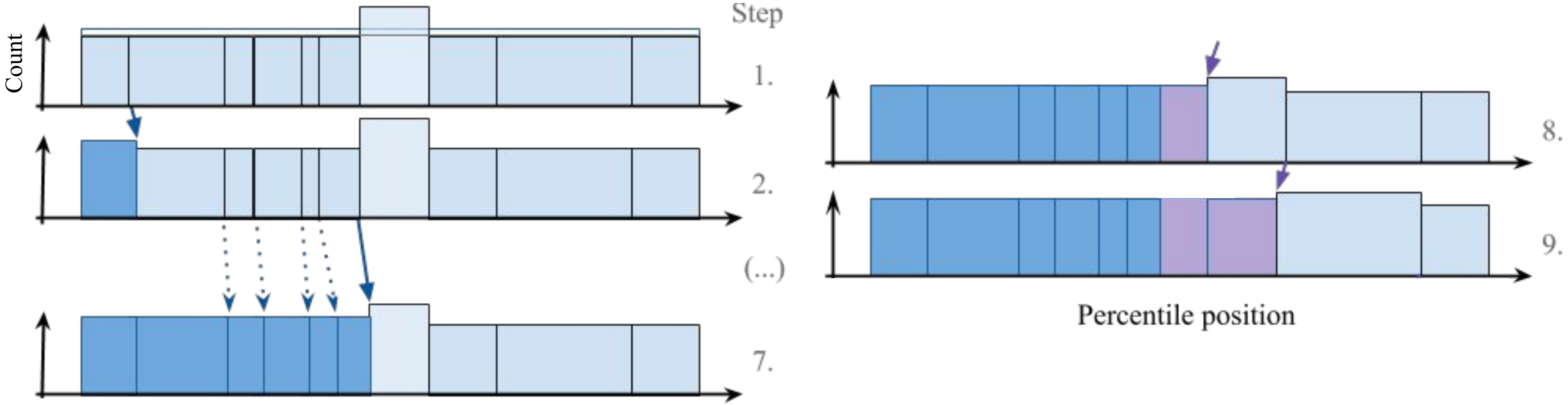
2b. The Threshold

SPEAR (Streaming Percentiles EstimAtoR)



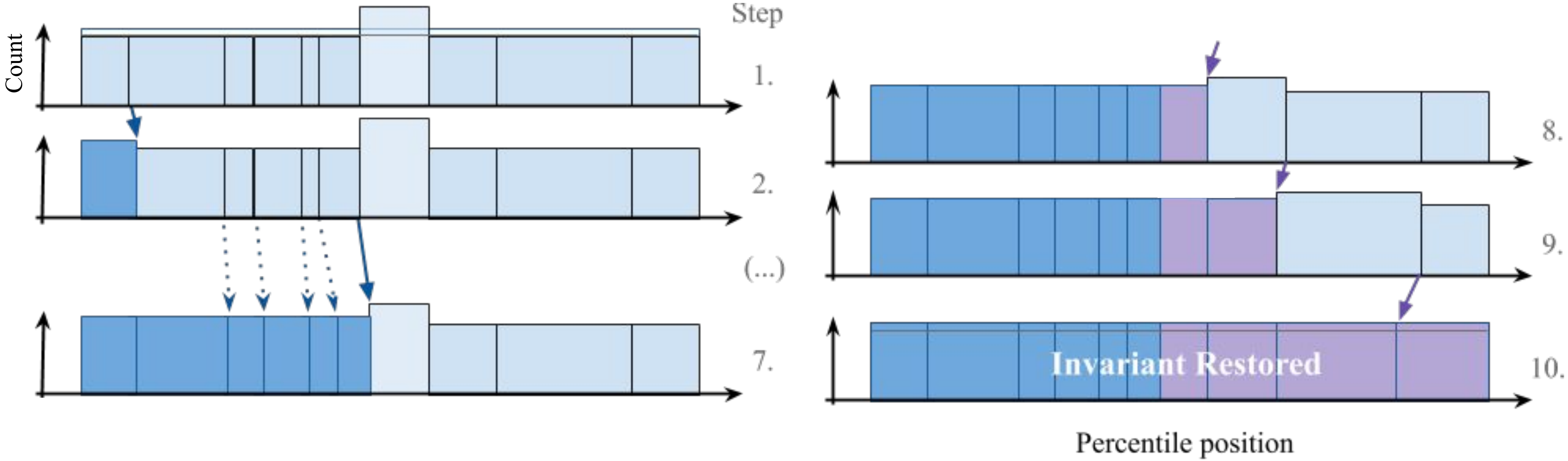
2b. The Threshold

SPEAR (Streaming Percentiles EstimAtoR)



2b. The Threshold

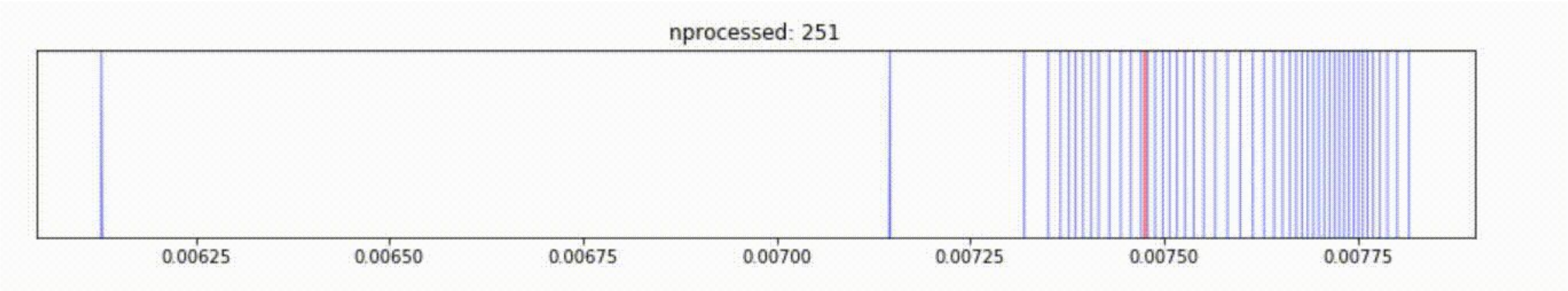
SPEAR (Streaming Percentiles EstimAtoR)



2b. The Threshold

AdaSPEAR (Adaptive SPEAR)

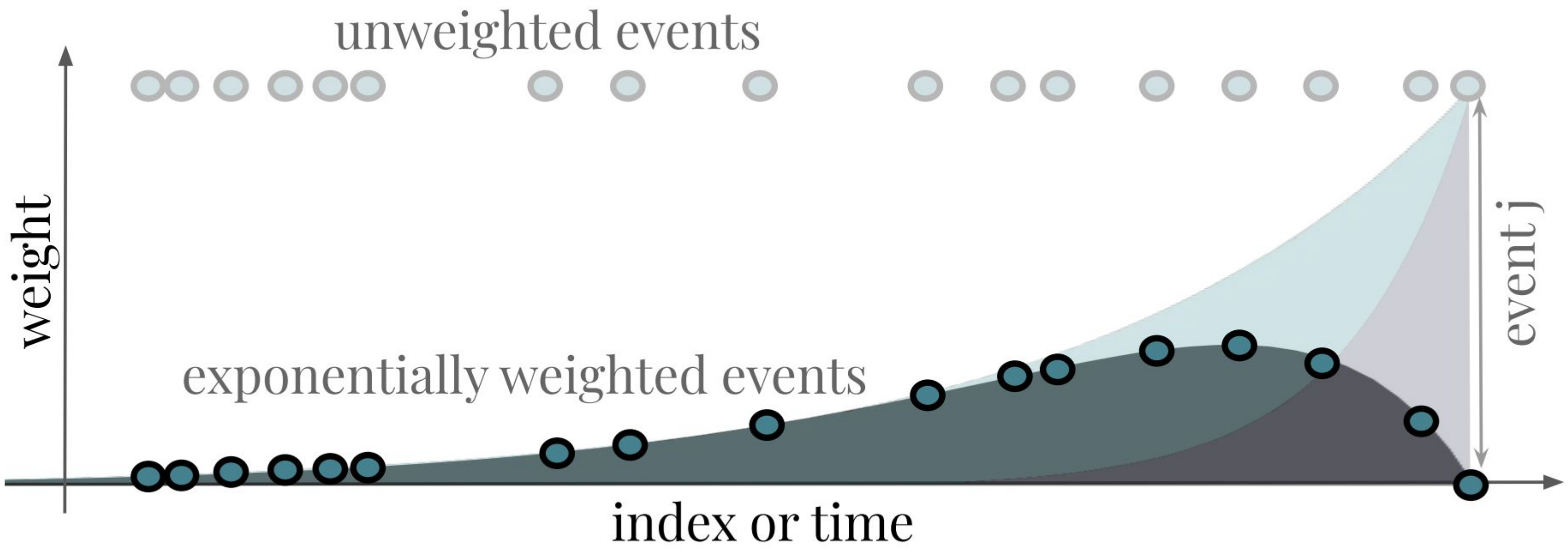
AdaSPEAR = SPEAR + suppress counts before bin expansion/contraction



Evolution of percentiles 0 to 100 (blue lines) in steps of 2. The red line indicates the most recent signal value.

2b. The Threshold

Delayed Adaptive Threshold

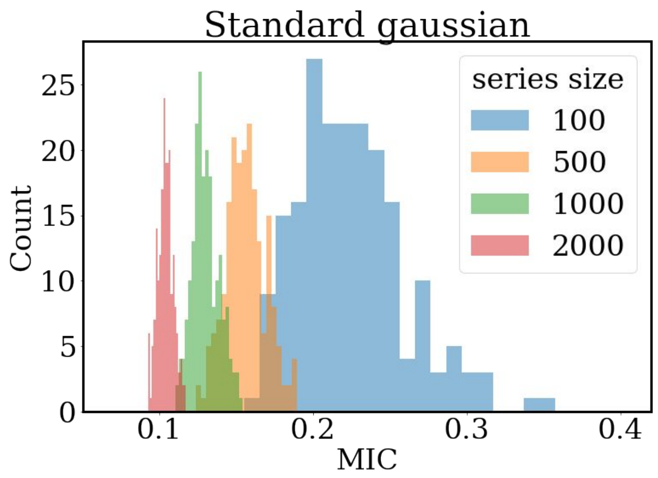


2c. Alarm report

Removal of Time Correlated Features

How?

- Preprocessing step for each feature that:
    a. Shuffles feature series values several times
    b. Computes correlation with time for each shuffle (MIC)
    c. Eliminates feature if MIC of original series in upper tail of distribution values with shuffle



2c. Alarm report

Removal of Time Correlated Features

Correlation of original series compared with random shuffles