TOWARDS A REAL-TIME UNSUPERVISED ESTIMATION OF PREDICTIVE MODEL DEGRADATION

Tania Cerquitelli, Stefano Proto,

Francesco Ventura, Daniele Apiletti,

Elena Baralis

BIRTE 2019

WHY CONCEPT DRIFT DETECTION?

From industrial production environments to smart cities, from network traffic classification to text mining

- data are collected in real-time
- the nature of data changes over time, due to the evolution of the phenomena

Predictive model performance usually degrades over time

- New incoming data can widely differ from the data distribution on which the model was trained
- Not all possible classes (labels) are effectively known at training time
- Real time predictions performed on new unseen data can be misleading or completely erroneous

STATE-OF-ART LIMITATIONS

Many techniques aim to be robust to concept drift

- They do not really detect concept drift and do not highlight drifting data

They require ground truth labels for drifting data to perform correctly

- They are applicable only in certain domains

They do not manage concept drift automatically and in real time

- They do not trigger predictive model retraining automatically only when necessary
- They are not thought to be scalable

Some approaches are not general purpose

- They are tailored to a specific use case

AUTOMATED CONCEPT DRIFT MANAGEMENT

Automatic triggering of the predictive model retraining only when necessary

Unsupervised approach

- It does not required the ground-truth labels for the newly classified samples

Explainable

- It produces description of the changes in the class-label data distributions motivating the model update

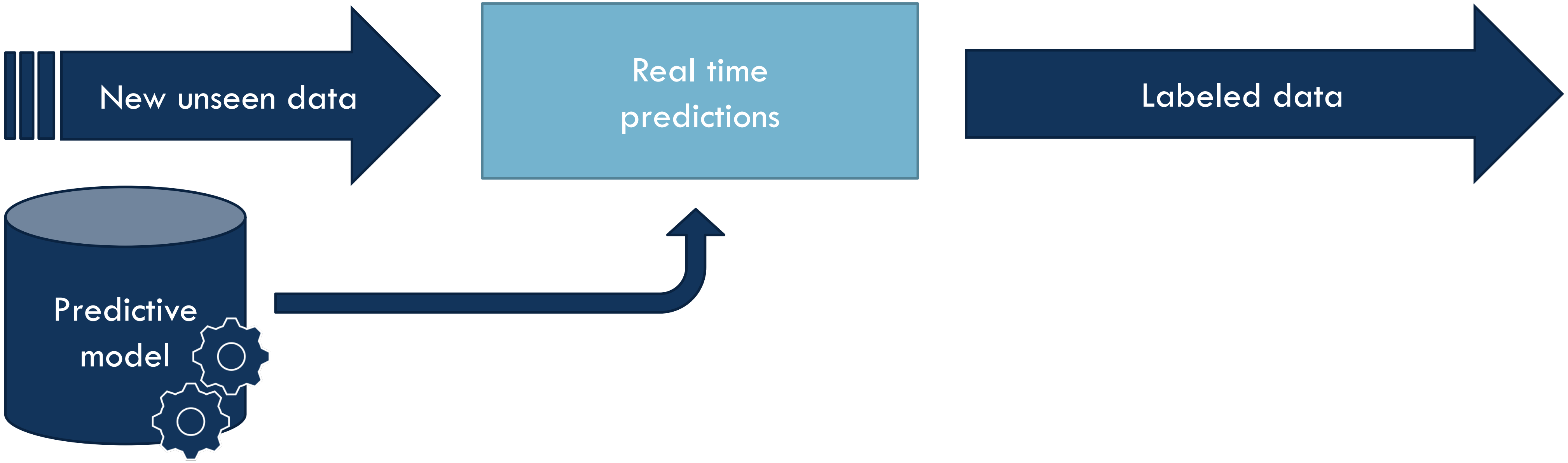
General purpose

- Not tailored to a specific use case or application domain, nor to a specific data type

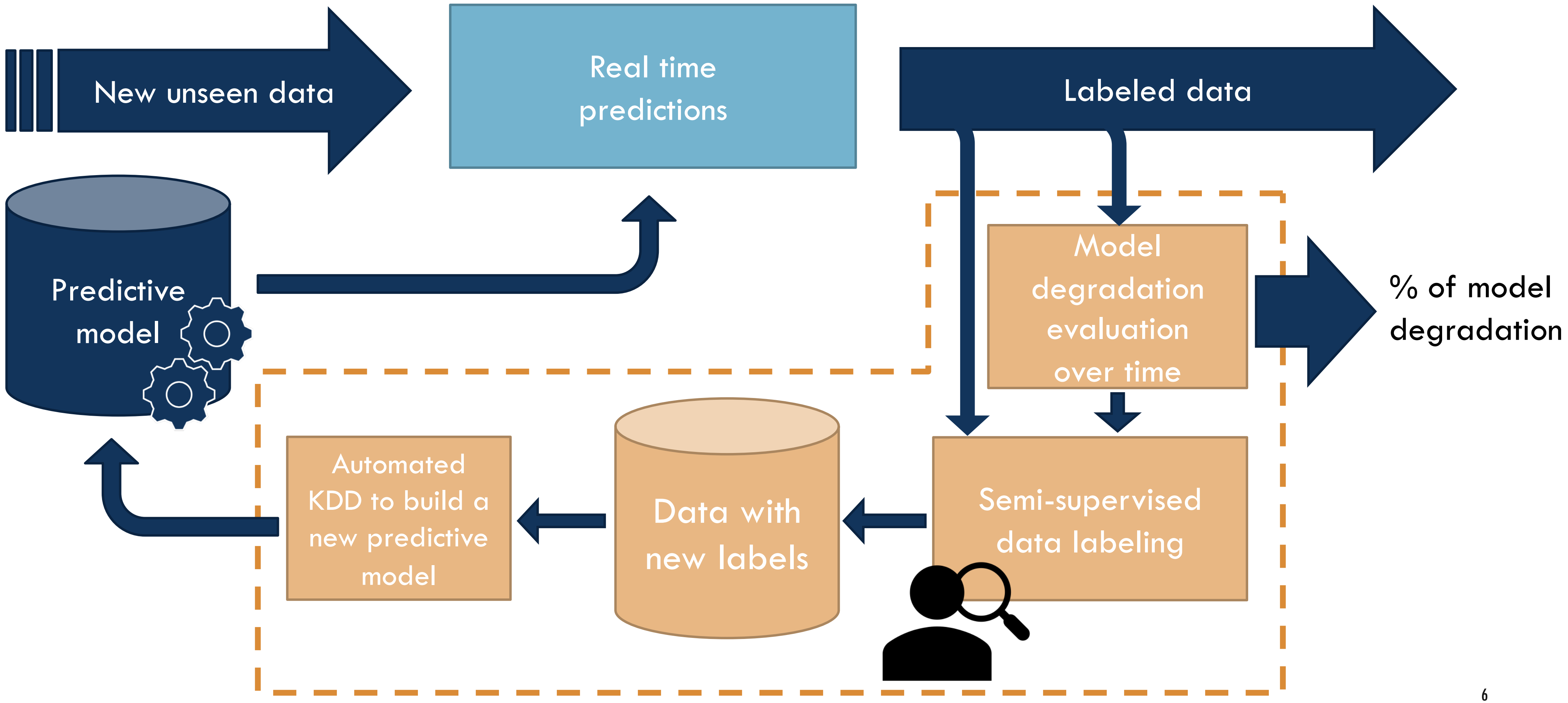
Real-time estimation

- Horizontally scalable for Big Data contexts and applicable in real-time environments
- Implemented on top of Apache Spark

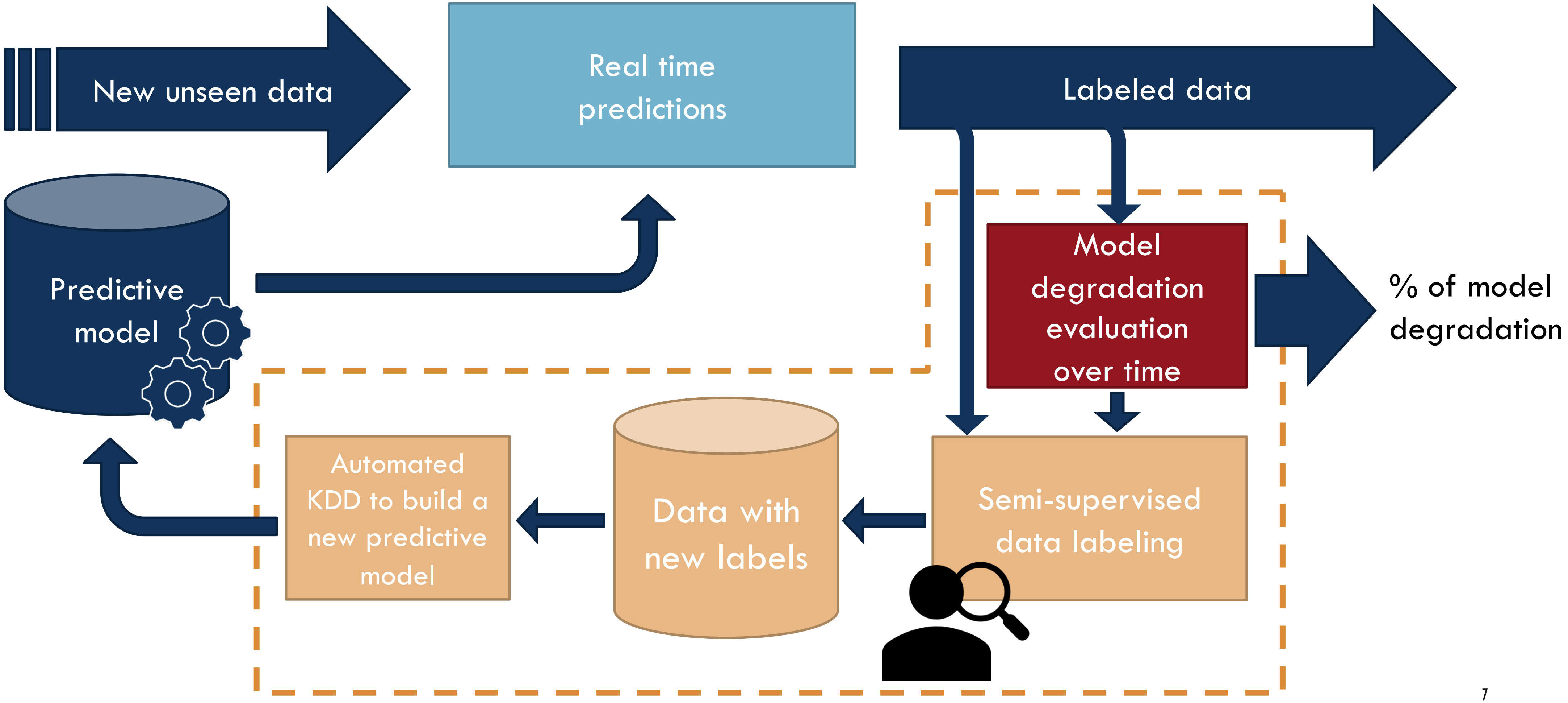
AUTOMATED CONCEPT DRIFT MANAGEMENT



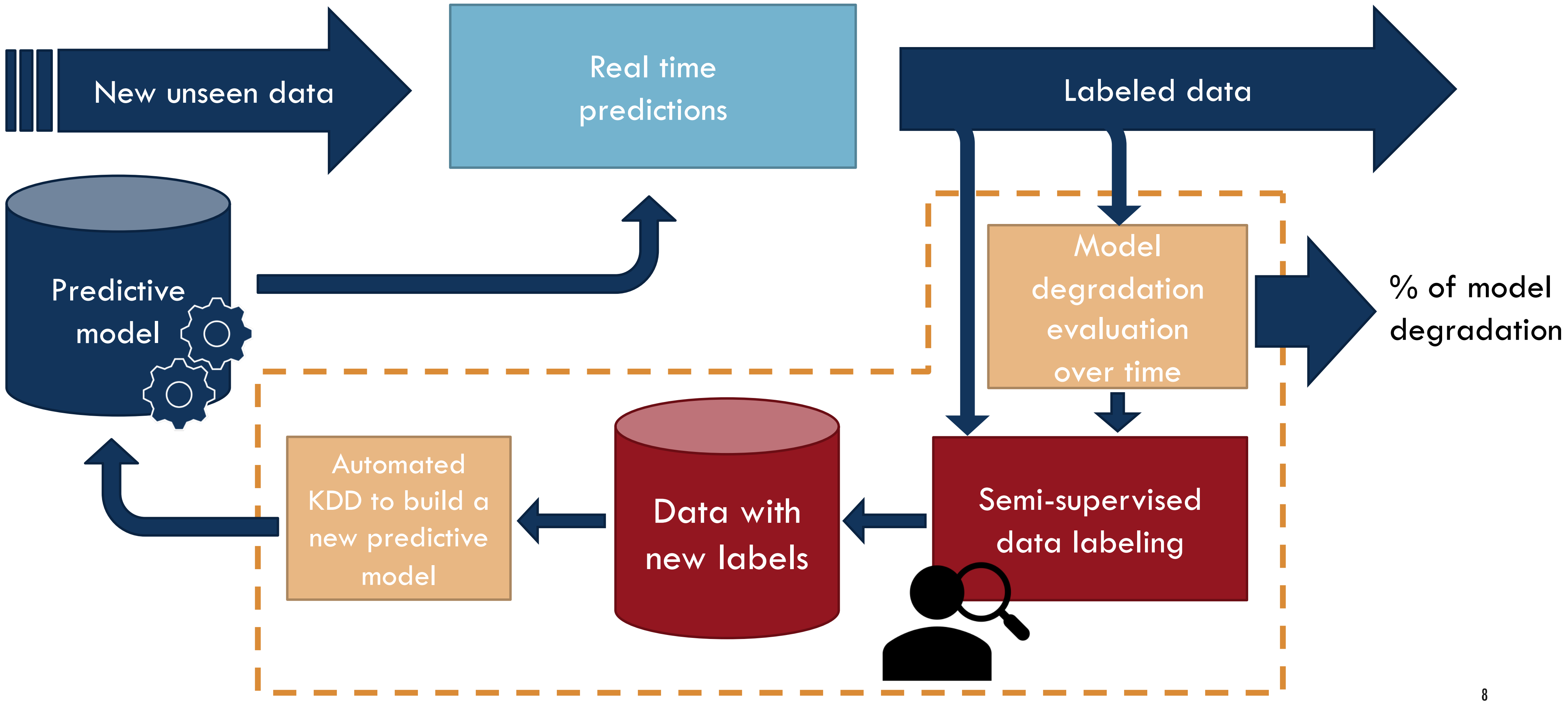
AUTOMATED CONCEPT DRIFT MANAGEMENT



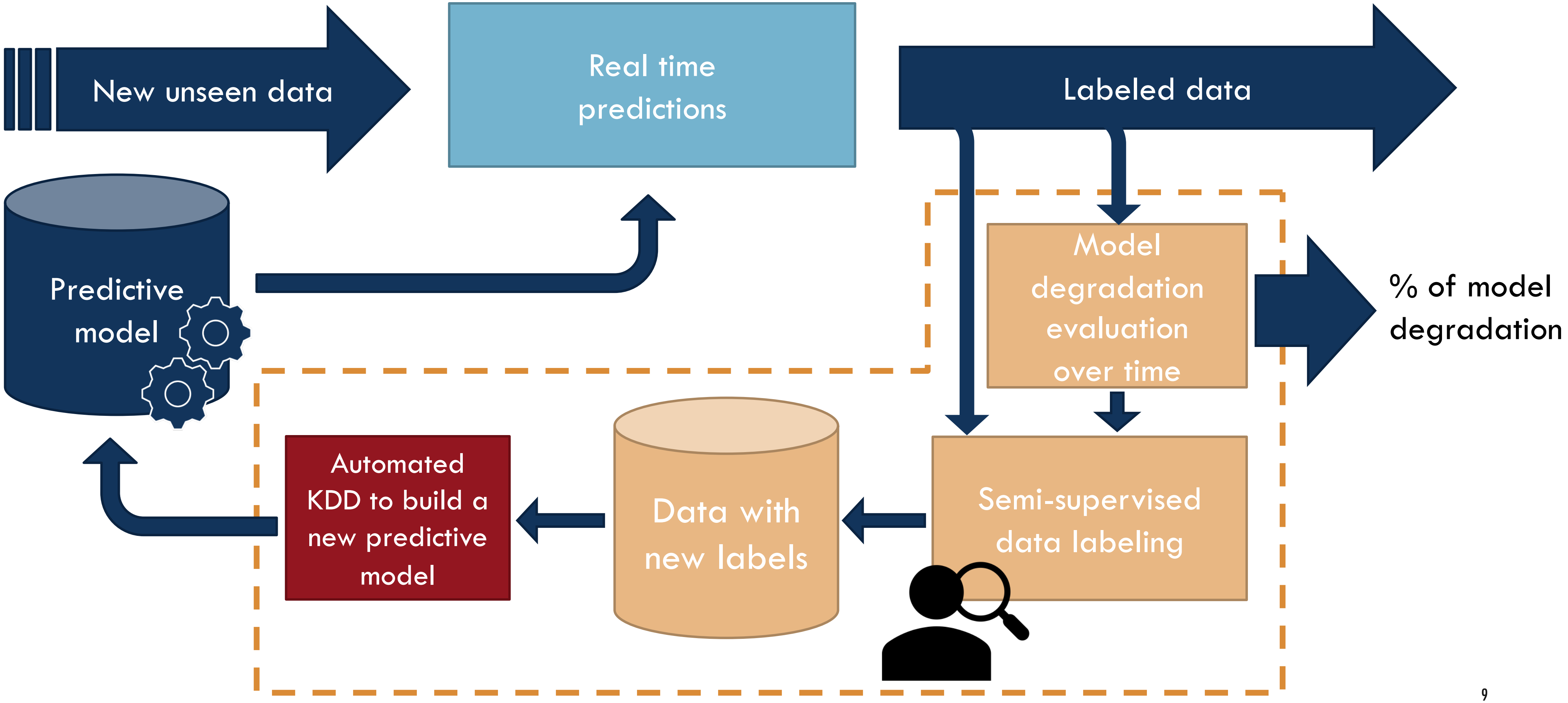
AUTOMATED CONCEPT DRIFT MANAGEMENT



AUTOMATED CONCEPT DRIFT MANAGEMENT



AUTOMATED CONCEPT DRIFT MANAGEMENT



MODEL DEGRADATION SELF-EVALUATION

METHODOLOGY

Given a pre-trained predictive model

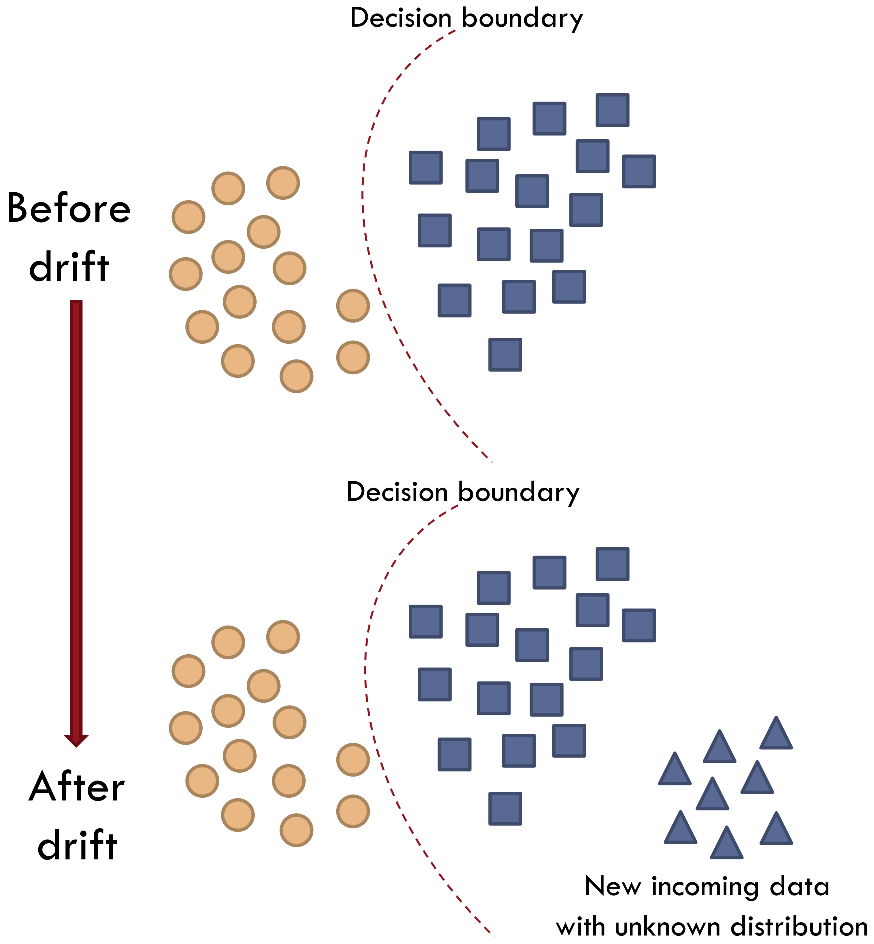
- Its knowledge is based on the information contained in the labeled train samples

We consider model performance degradation between

- Data used to train the classification model
- New incoming unlabeled data

Algorithm main idea

- given a dataset of points divided in classes
- Evaluate the **intra-class cohesion** and **inter-class separation**
- **Before** and **After** the prediction of unseen data
- Compute the degradation of the predictive model.



| ***Color*** is the class label assigned by the classifier
| ***Shape*** is the ground-truth class label

MODEL DEGRADATION SELF-EVALUATION

METHODOLOGY

The self-assessment algorithm exploits *unsupervised quality metrics* to *evaluate the predictive model degradation*

The algorithm exploits the scalable **Descriptor Silhouette** index (DS)

- Other unsupervised metrics can be used

The **Model Degradation** is obtained computing the MAAPE error between

- Descriptor Silhouette curve computed at the end of the model training with training data at time t_0
- Descriptor Silhouette curve computed with training data and new labeled data until time t

Degradation is computed separately for each class

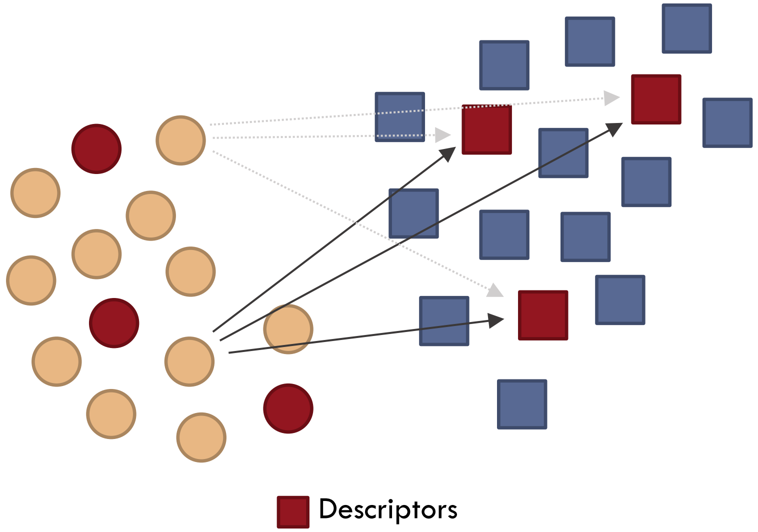
METHODOLOGY
DESCRIPTOR SILHOUETTE INDEX[1]

The geometrical shape of a group of points is described with a low number of **Descriptors**

The DESCRIPTOR SILHOUETTE[1] applies the same definition of Silhouette

- between all the points in the dataset and the descriptors

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}}$$



1. Francesco Ventura, Stefano Proto, Daniele Apiletti, Tania Cerquitelli, Simone Panicucci, Elena Baralis, Enrico Macii, and Alberto Macii. 2019. **A new unsupervised predictive-model self-assessment approach that SCALEs**. In 2019 IEEE International Congress on Big Data (BigData Congress). IEEE, 144–148. https://doi.org/10.1109/BigDataCongress.2019.00033

METHODOLOGY
MODEL DEGRADATION

$$DEG(c,t) = \alpha * \mathrm{MAAPE}(Sil_{t_0}, Sil_t) * \frac{N_c}{N}$$

$$\alpha = \begin{cases} 1 & if: \ \overline{Sil_{t_0}} \geq \overline{Sil_t} \\ -1 & if: \ \overline{Sil_{t_0}} < \overline{Sil_t} \end{cases}$$

$DEG(c,t)$ → Model Degradation for class c at time t

$MAAPE(a,b)$ → Mean Absolute Arctangent Percentage Error

Sil_{t_0} → Descriptor Silhouette at training time

Sil_t → Descriptor Silhouette at training time + labeled data until time t

$\frac{N_c}{N}$ → Ratio between #points belonging to class c and total number of points

α → Coefficient that is positive or negative according to the comparisons of average silhouettes at time t_0 and t

EXPERIMENTAL GOALS

Prove the effectiveness of **model degradation self-evaluation** over time.

Show the performances of the **Descriptor Silhouette**

EXPERIMENTAL CONTEXT 1
MODEL DEGRADATION SELF-EVALUATION

2datasets

Dataset D1

- Synthetic dataset created with the scikit-learn Python library
- 800,000 records
- 4 normally distributed classes (200,000 for each class)
- 10 features

Dataset D2

- Real-world dataset containing Wikipedia articles
- 3,000 records
- 3 classes: food-drink, literature and mathematics – 1000 records for each class
- 100 features obtained through Doc2Vec document embedding

EXPERIMENTAL CONTEXT 1
MODEL DEGRADATION SELF-EVALUATION

Random Forest classifier has been used as predictive model.

- 3-fold cross-validation
- average f-measure of the predictive model
  - 0.964 for dataset D1
  - 0.934 for dataset D2.

The training set consists of a stratified sample over classes 0 and 1 with 60% of records in each class.

The remaining part of the dataset is used as test set to assess model degradation

- 40% of classes 0 and 1 and whole class 2

EXPERIMENTAL RESULTS

MODEL DEGRADATION SELF-EVALUATION - 1

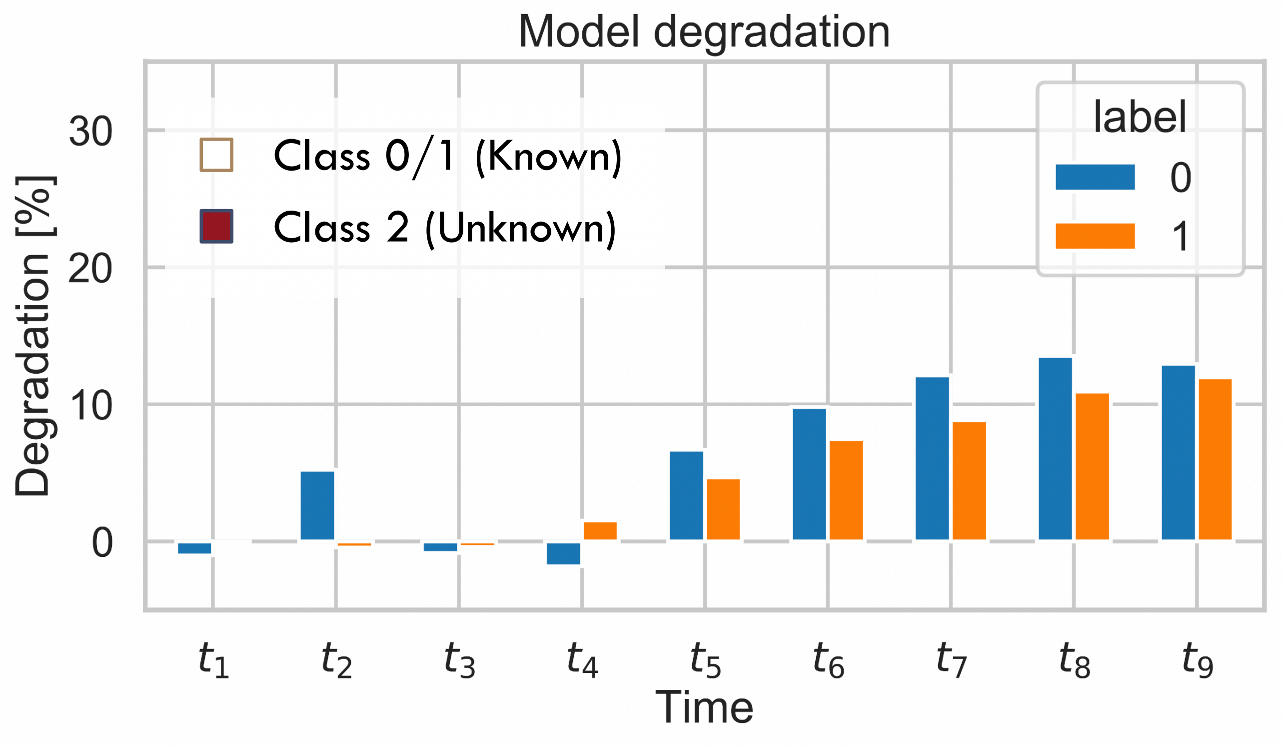
Dataset D1

Training on 60% of

- Classes 0 and 1

Test degradation on

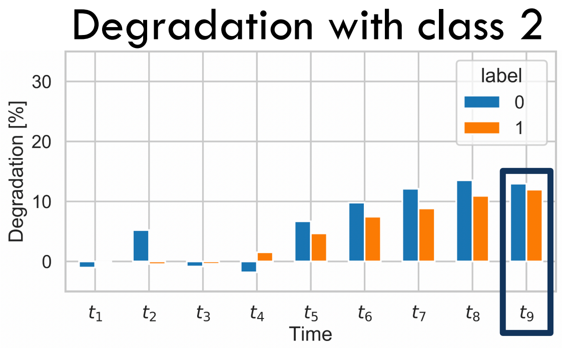
- 40% classes 0 and 1
- Whole class 2

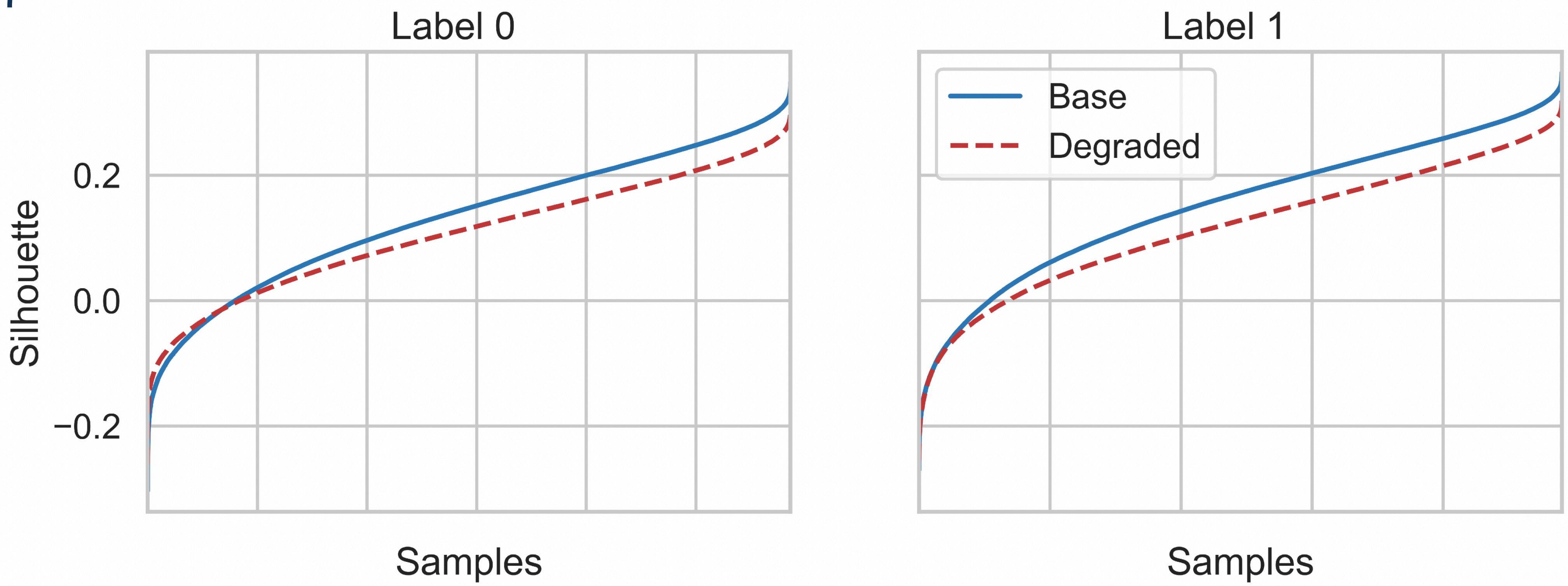


Percentage of drifting records	0%	0%	0%	0%	9%	16.7%	23.1%	28.6%	33.3%
Number of drifting records	0	0	0	0	40K	80K	120K	160K	200K

New incoming data

EXPERIMENTAL RESULTS
MODEL DEGRADATION SELF-EVALUATION - 1





Dataset D1. Baseline DS curve at training time, and degraded DS curve at time t9

EXPERIMENTAL RESULTS

MODEL DEGRADATION SELF-EVALUATION - 2

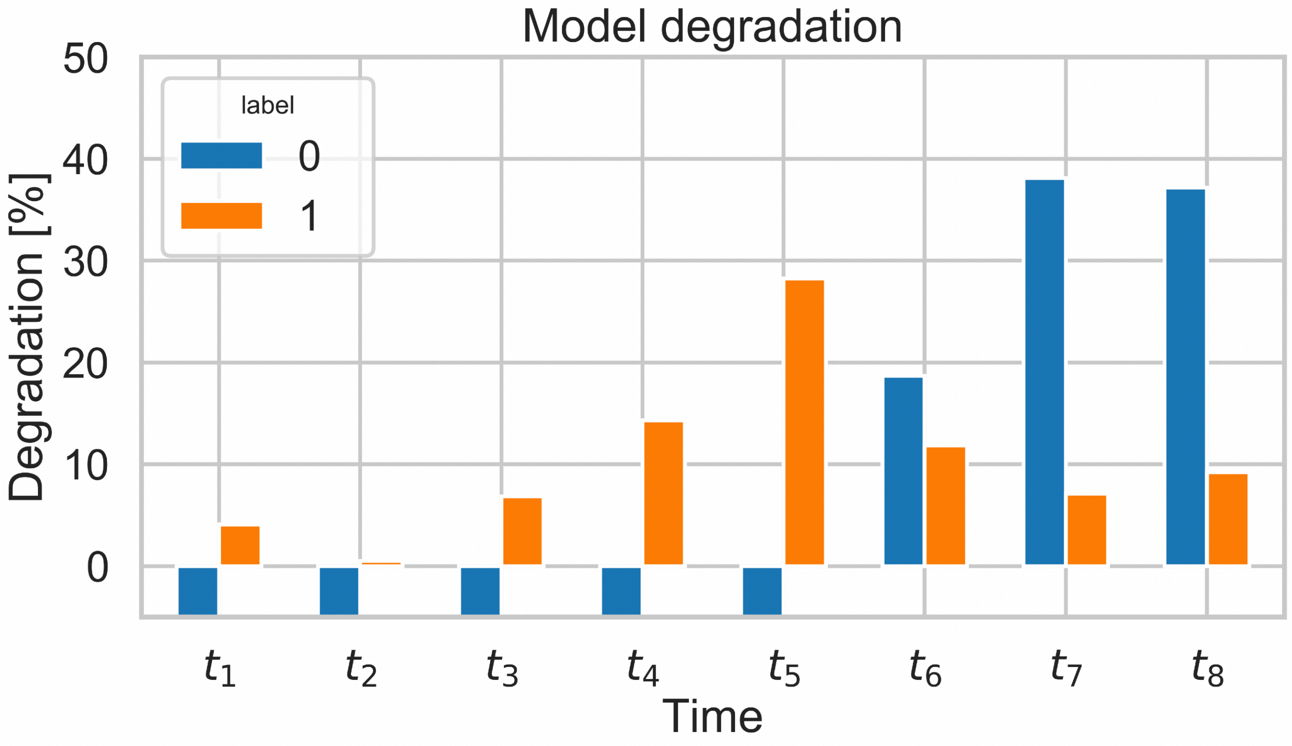
Dataset D2 - Wikipedia

Training on 60% of

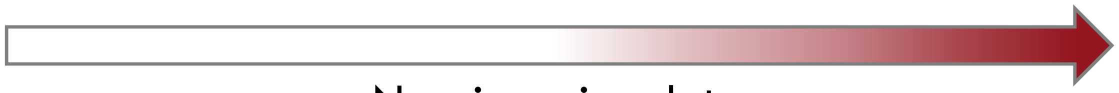
- Class 0 (food-drink)
- Class 1 (literature)

Test degradation on

- 40% classes 0 and 1
- Class 2 (mathematics)



Percentage of drifting records	0%	0%	0%	0%	9%	16.7%	23.1%	28.6%
Number of drifting records	0	0	0	0	200	400	600	800



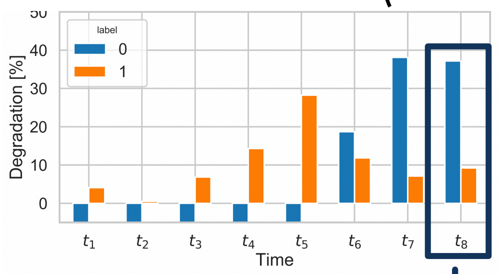
New incoming data

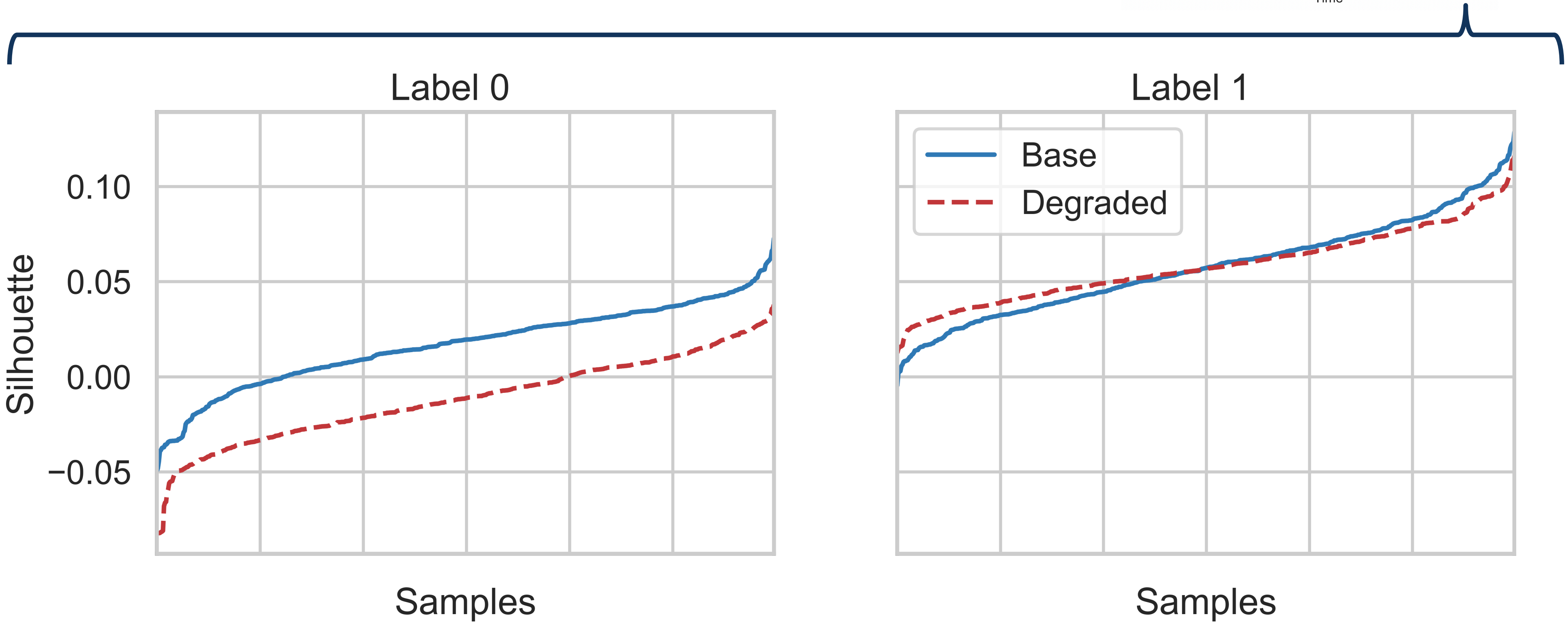
- Class 0/1 (Known)
- Class 2 (Unknown)

EXPERIMENTAL RESULTS

MODEL DEGRADATION SELF-EVALUATION - 2

Degradation with class 2 (mathematics)





Dataset D2. Baseline DS curve at training time, and degraded DS curve at time t8

EXPERIMENTAL CONTEXT 2
DESCRIPTOR SILHOUETTE PERFORMANCE

Synthetic dataset

- 10M records
- 10 features
- 3 classes
- Normal distribution

200 descriptors per class

6 sub-datasets

- 10k, 50K, 100K, 500K, 1M, 10M

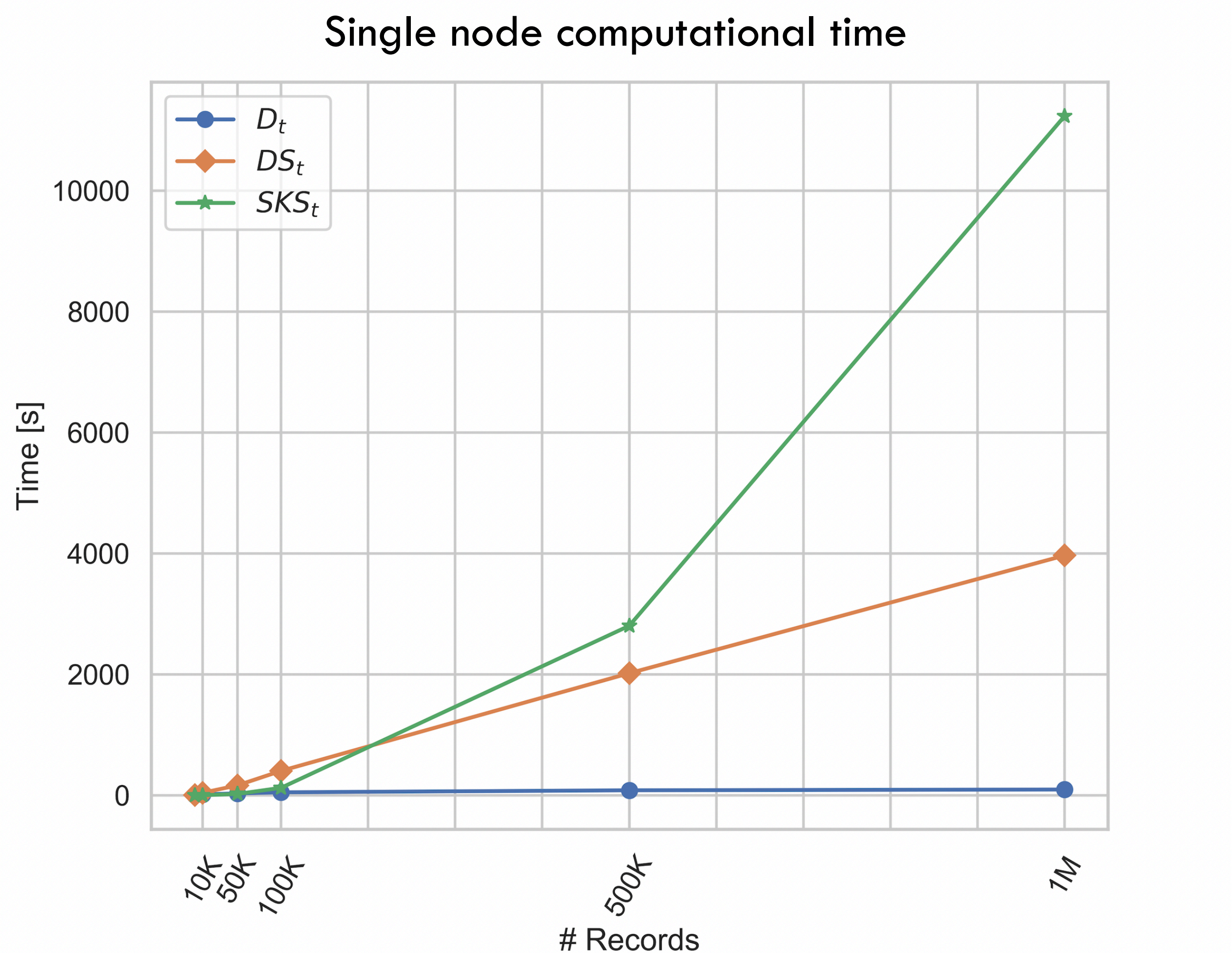
Single node configuration

- Intel i7 8-core server
- 32GB of memory

Multi node configuration

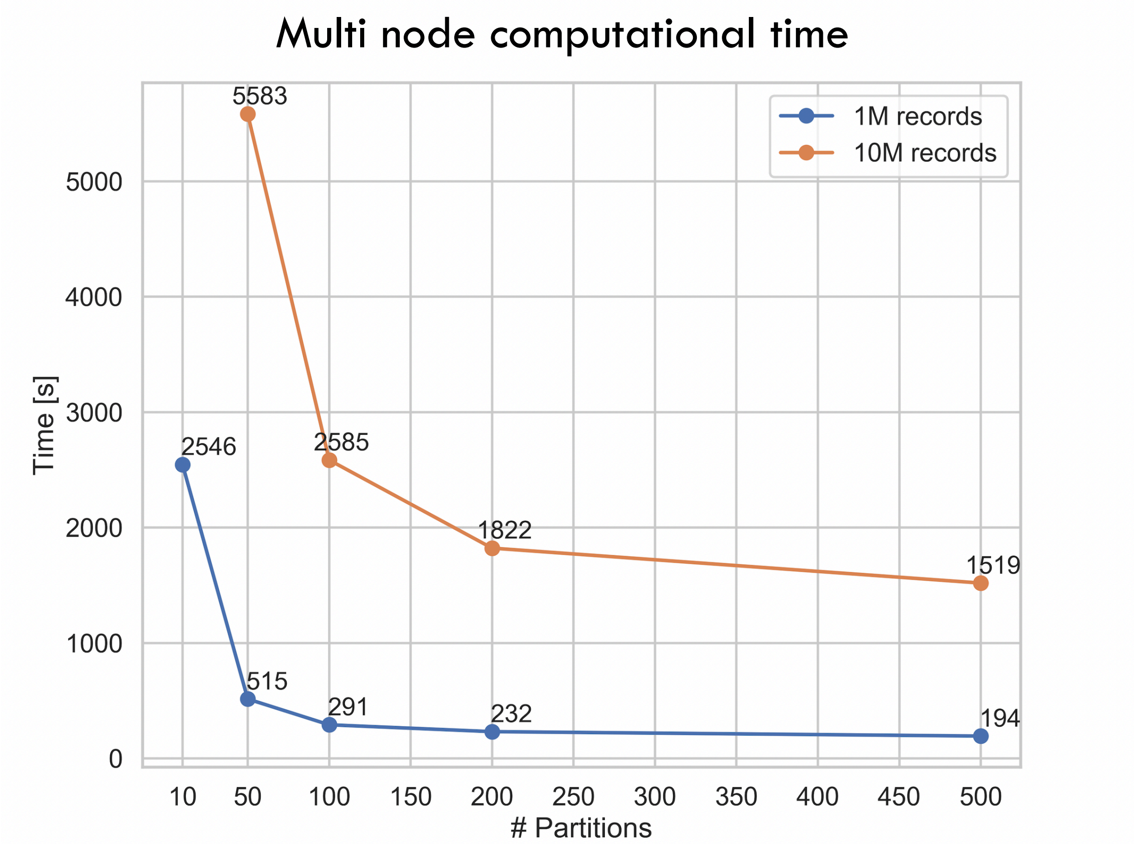
- 50 virtual nodes
- 2 cores
- 512MB of memory
- running on top of the BigData@Polito cluster (https://smartdata.polito.it/computing-facilities/)

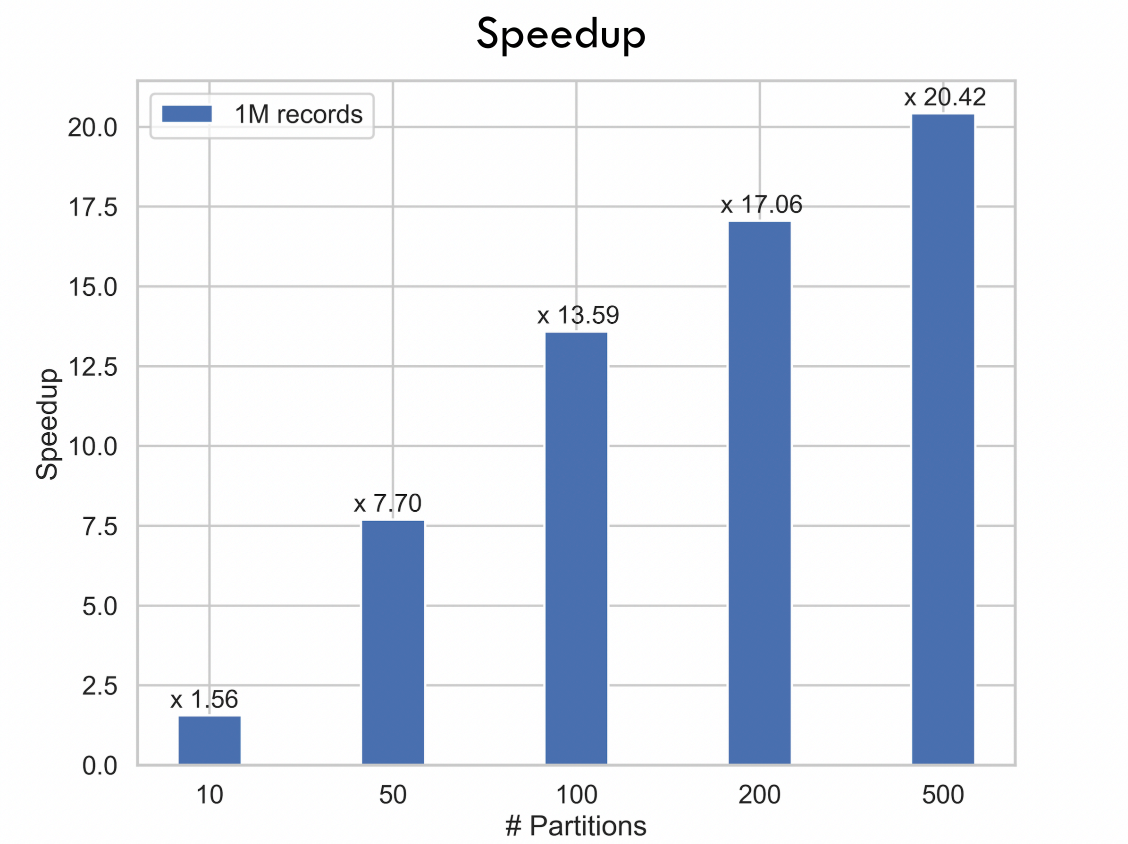
EXPERIMENTAL RESULTS
DESCRIPTOR SILHOUETTE PERFORMANCE – SINGLE NODE



EXPERIMENTAL RESULTS

DESCRIPTOR SILHOUETTE PERFORMANCE – MULTI NODE





When data is distributed in 500 partitions over the 50 nodes, the Descriptor Silhouette index requires:

- 25 mins for 10M records
- 3 mins for 1M records

CONCLUSIONS & FUTURE WORK

Automated concept drift management with a new estimation strategy for model degradation

- In soft real-time
- Exploiting an unsupervised strategy
- General purpose

Promising experimental results on two datasets

Future directions include

1. alternative unsupervised metrics besides the Silhouette index
2. improvement of self-evaluation triggering mechanism, currently set as a percentage of new data
3. further experiments to assess the generality and the real-time performance

Francesco Ventura

Politecnico di Torino (Italy)

francesco.ventura@polito.it

www.linkedin.com/in/f-ventura

THANKS

Tania Cerquitelli, Stefano Proto,

Francesco Ventura, Daniele Apiletti,

Elena Baralis

This work has been partially funded by the SmartData@Polito center for Data Science and Big Data technologies, Politecnico di Torino, Italy.

BIRTE 2019