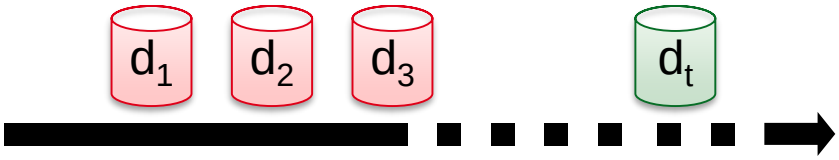
FROM RESEARCH TO INDUSTRY

cea tech

list

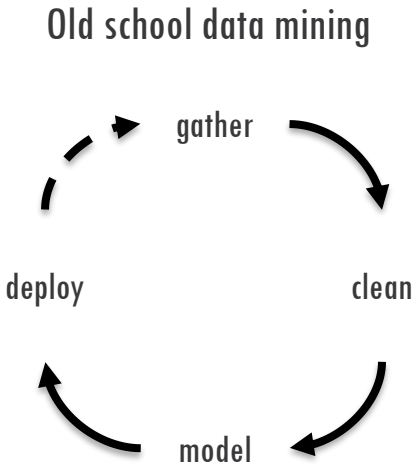


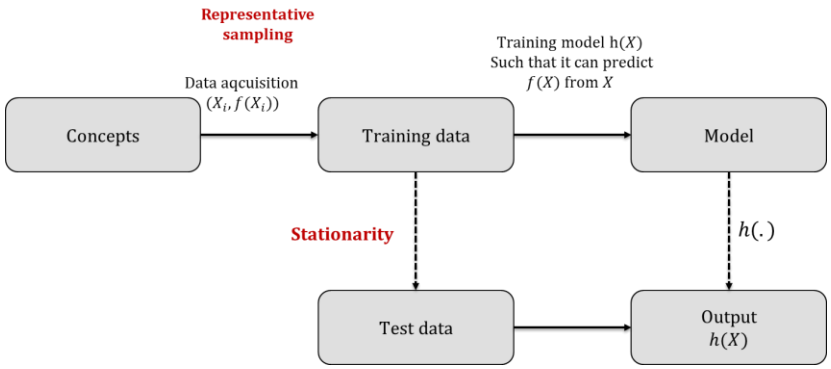
STREAMER

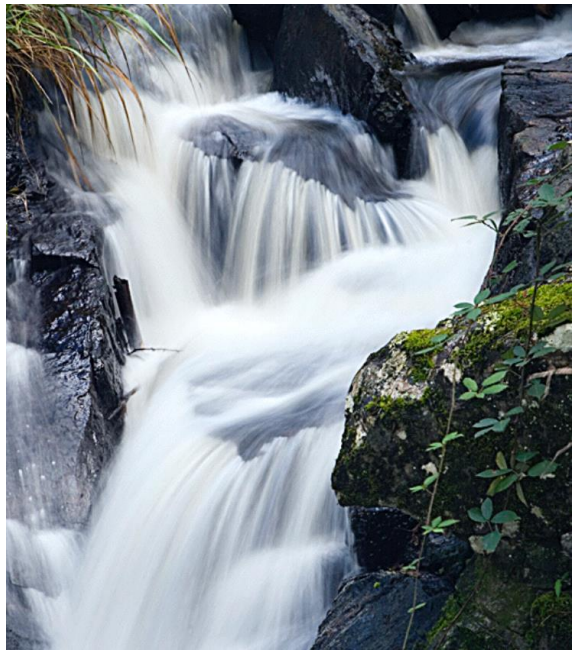
A POWERFUL AND OPEN-SOURCE FRAMEWORK FOR CONTINUOUS LEARNING IN DATA STREAMS

Sandra GARCIA RODRIGUEZ

AI PARADIGMS: FROM CLASSICAL TO STREAM LEARNING







Paradigm shift

2 consequences

- Get a model while data is still incoming
- Non-stationarity

DATA STREAMS

- **Definition**
  - Infinite streams of data integrated from both live and historical sources
  - Stream: [d_0,d_1… d_n]

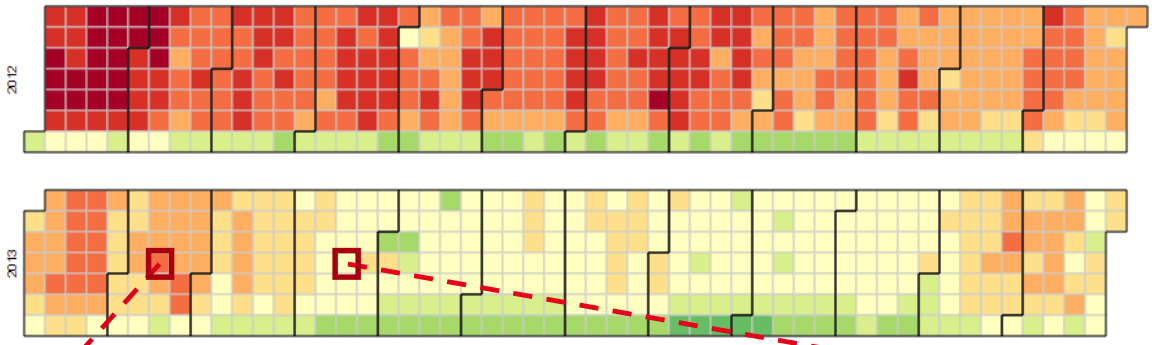
  - They contain entities [and labels] <*e,L*>
  - Each entity *e* is linked to *1..n* instances (pieces of information)

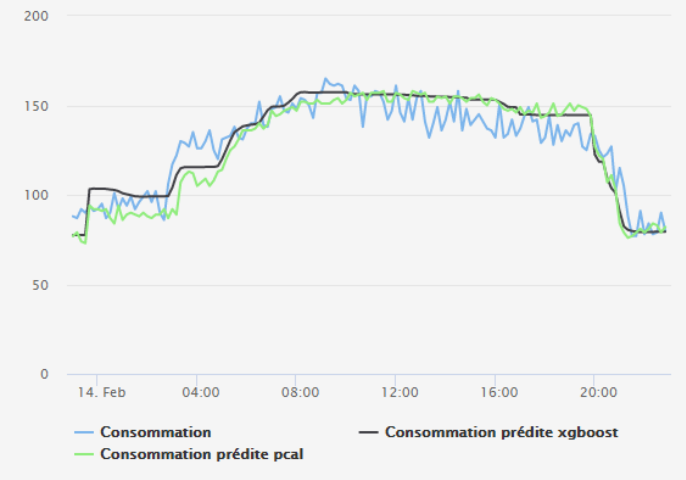
	Batch processing	**Stream processing**
Data scope	Queries or processing over all or most of the data in the dataset.	Queries or processing over data within a rolling time window, or on just the most recent data record.
Data size	Large batches of data.	Individual records or micro batches consisting of a few records.
Performance	Latencies in minutes to hours.	Requires latency in the order of seconds or milliseconds.
Analyses	Complex analytics.	Simple response functions, aggregates, and rolling metrics.

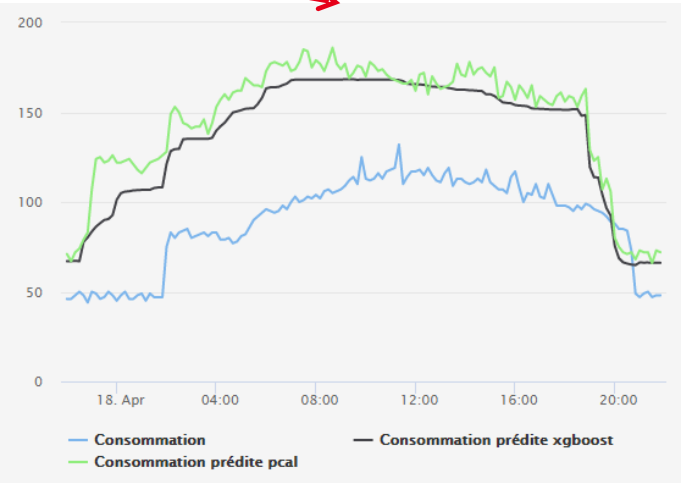
- **DS processing tools** (as STREAMER)
  - algorithms applied to DS that can learn incrementally

IMPORTANCE OF ADAPTATIVE MODELS

Non-stationarity in electricity consumptions







DE LA RECHERCHE À L'INDUSTRIE

cea

www.cea.fr

WHAT IS STREAMER?

INSTITUT CARNOT CEA LIST

digiteo

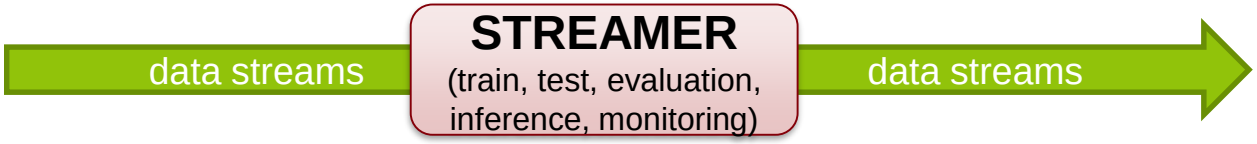
list

INTRODUCTION

Data Streams: data that arrive continuously but cannot be stored, it needs to be processed in real time.

- **What is it?** framework for integrating and testing machine learning algorithms in realistic streaming operational contexts
  - in multiple programming languages
  - some already available

- **Users** can abstract from streams implementation (transparent for them) to only focus on the algorithms



- Implementation of the whole pipe (or just part of it) of data stream ingestion, processing, knowledge extraction and visualization

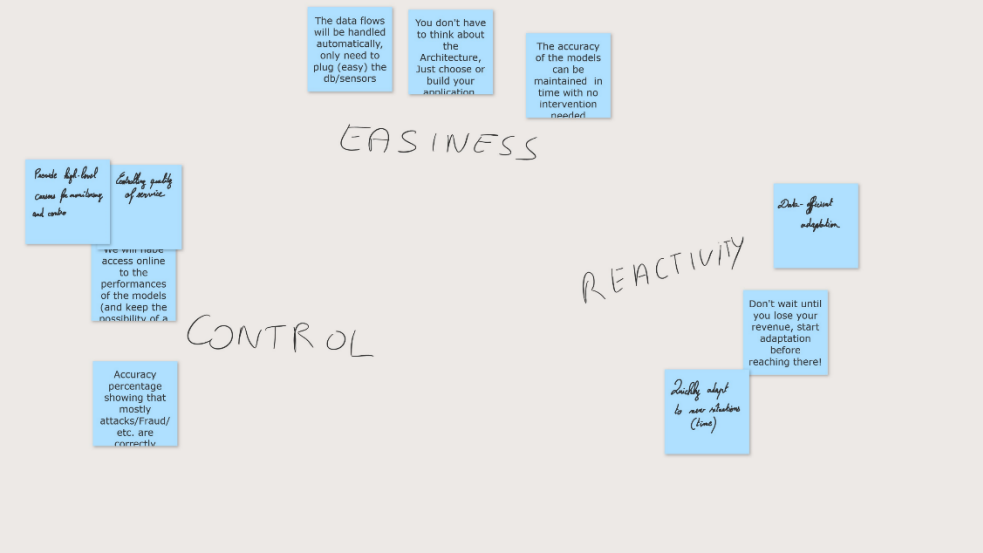
- **Features**:
  - open source
  - flexible & scalable
  - cross-platform
  - logs
  - visualization

OBJECTIVES

- **Control - machine learning experimenter**
  - Allows users to easily integrate and test machine learning algorithms into realistic streaming operational context
  - User must be able to define the context and learning configuration
  - Provides several functionalities (streaming scenario simulation, train, test, evaluation, models inference, monitoring…)

- **Easiness**
  - Easy installation and deployment
  - Visualization capabilities

- **Reactivity**
  - Process and incrementally learn from continuous data streams



TARGETED USERS AND FUNCTIONALITIES

Sandbox for evolving AI design in « real-life conditions »

Expertise level

Data owner

GUI-assisted design

Synopsis construction
Sampling

Forecasting

Clustering

Dimensionality reduction

STREAMER

Classification

Frequent pattern mining

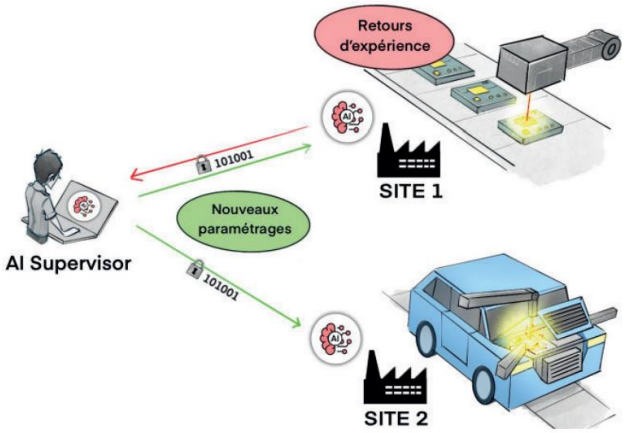
Monitoring
Change detection

Data scientist

Config files

Current STREAMER

Evolving AI in real-life operations



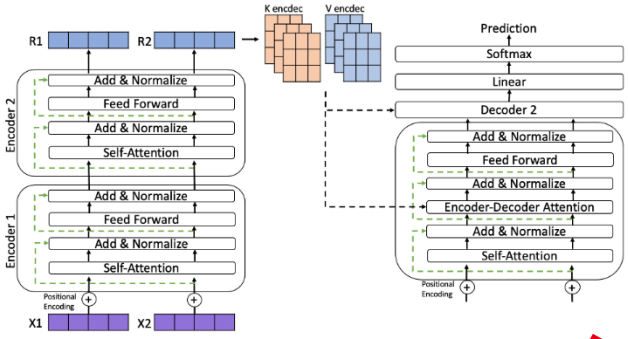
User
Automation

Governor
Empowerment

Regulator
Auditability

Robust and secure deployment

STREAM LEARNING FUNCTIONALITIES



- Clustering.java
- HoeffdingTreeClassifier.java
- MLalgorithms.java
- NeuralNetwork.java
- SVM.java
- XGBoost.java

Custom algorithms

Stream learning functionalities

Existing renowned frameworks

scikit learn

scikit-multiflow

dmlc XGBoost

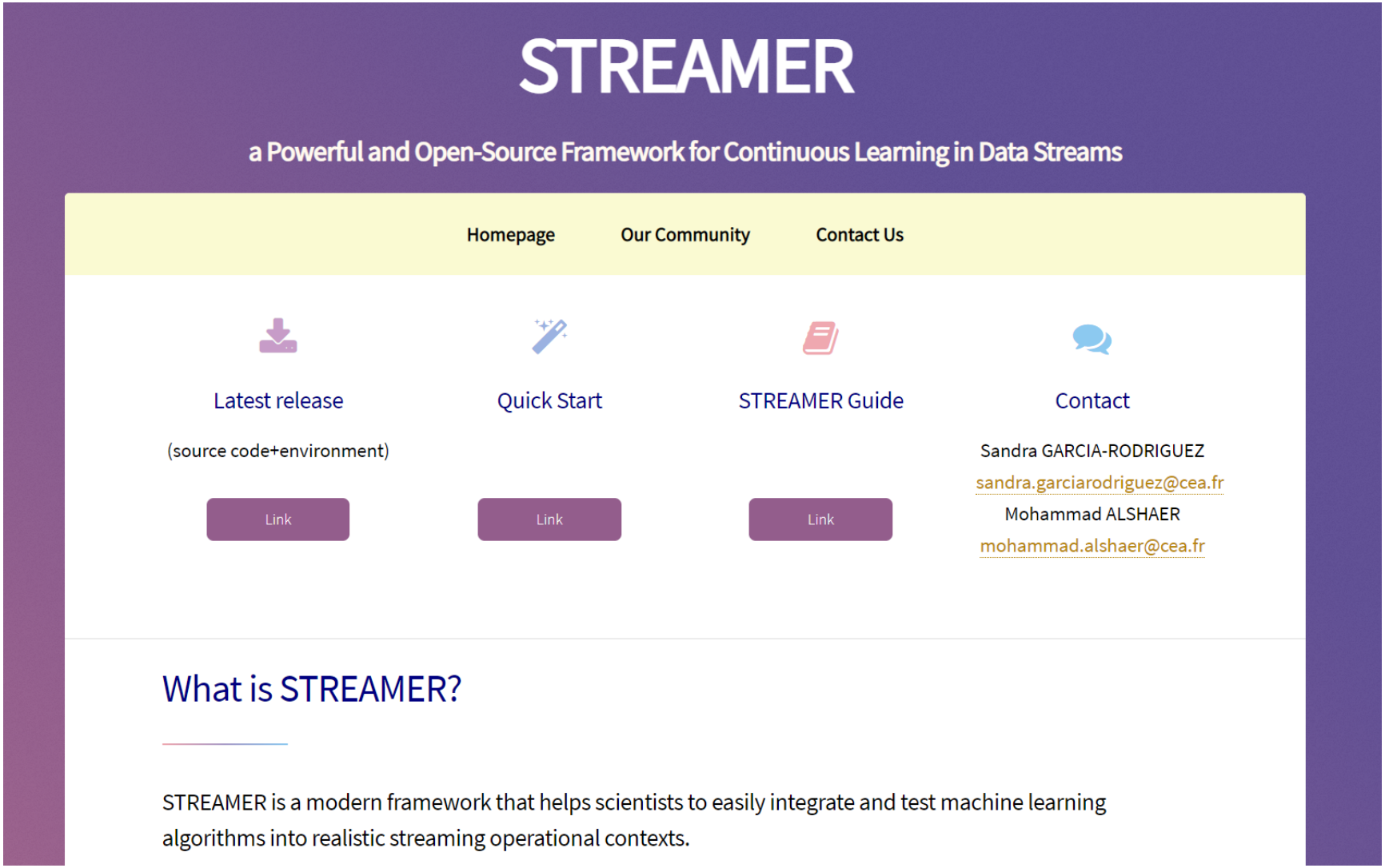
TensorFlow 2.0

Custom metrics

- MultiClassClassificationMetric.java
- MultiConfusionMatrix.java
- Pair.java
- PrecisionBinaryClassificationMetric.java
- PrecisionMultiClassClassificationMetric.java
- QuantileSummary.java
- SensitivityBinaryClassificationMetric.java
- SensitivityMetric.java
- SensitivityMultiClassClassificationMetric.java
- AccuracyBinaryClassificationMetric.java
- AccuracyMetric.java
- AccuracyMultiClassClassificationMetric.java
- BinaryClassificationMetric.java
- F1scoreBinaryClassificationMetric.java
- F1scoreMultiClassClassificationMetric.java

STREAMER IN DETAILS

- **The code is published in open-source on GitHub**
- **Official Website:** *https://streamer-framework.github.io/*



DE LA RECHERCHE À L'INDUSTRIE

cea

www.cea.fr

STREAMER
MORE IN DETAIL

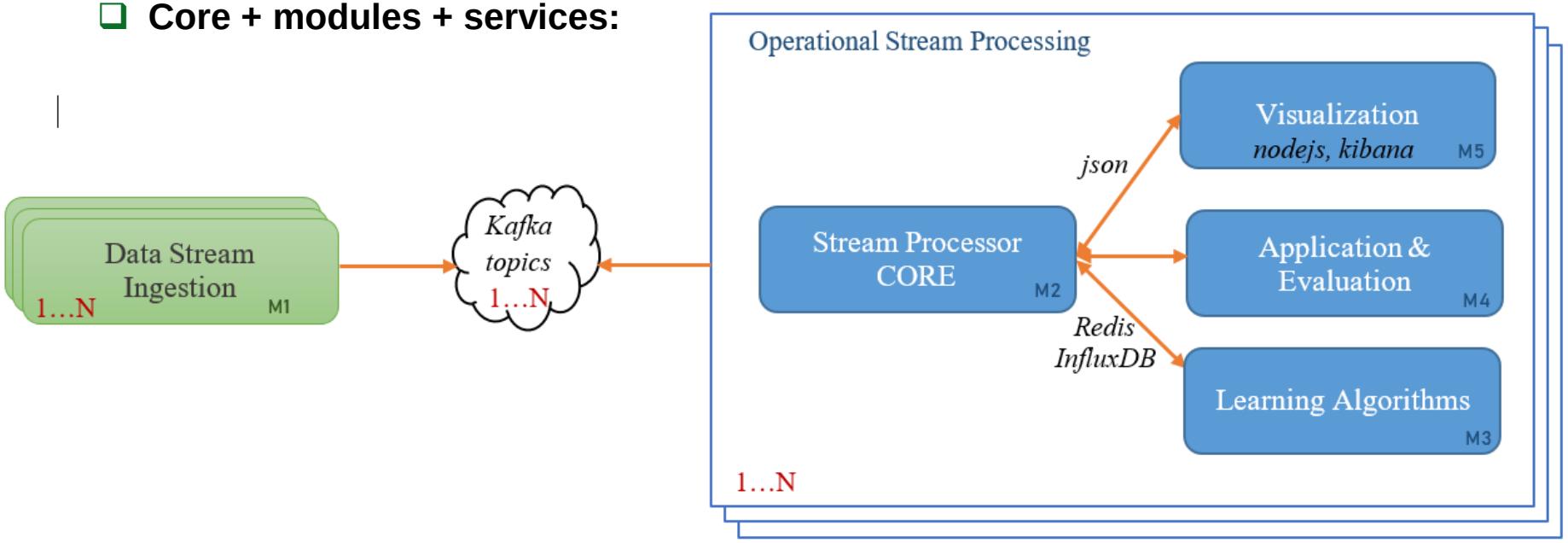
INSTITUT CARNOT CEA LIST

digiteo

list

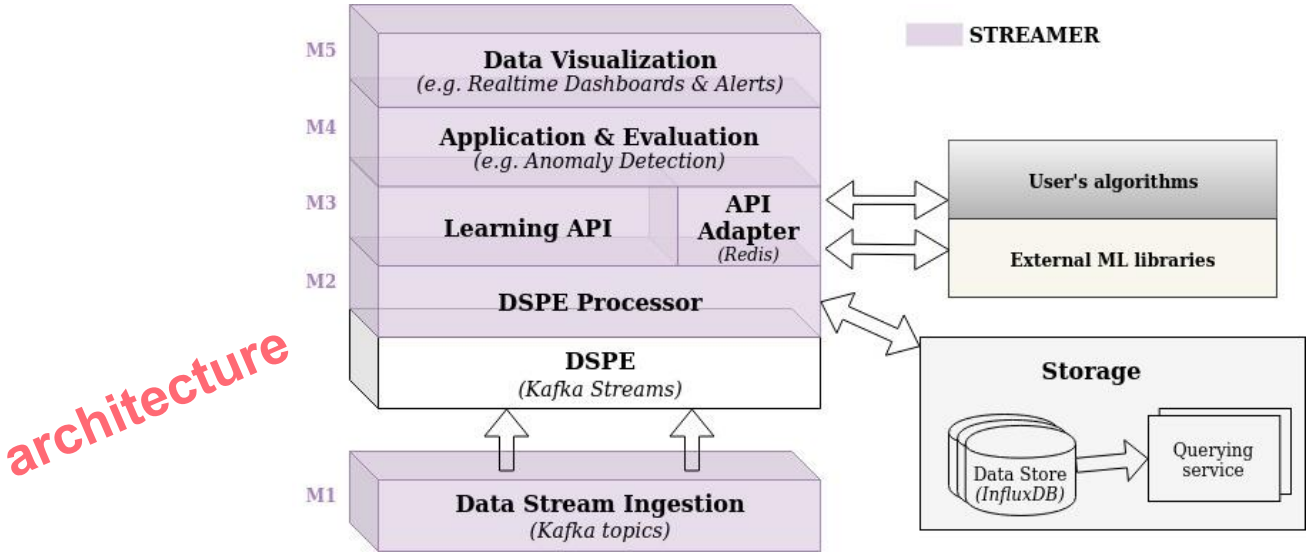
- **License** (use and distribution): terms by GNU GENERAL PUBLIC LICENSE Version 3 (https://www.gnu.org/licenses/gpl-3.0.html)
- Easy **installation**:
  - Docker environment
  - Manual choice (advanced mode)

- **Core + modules + services:**

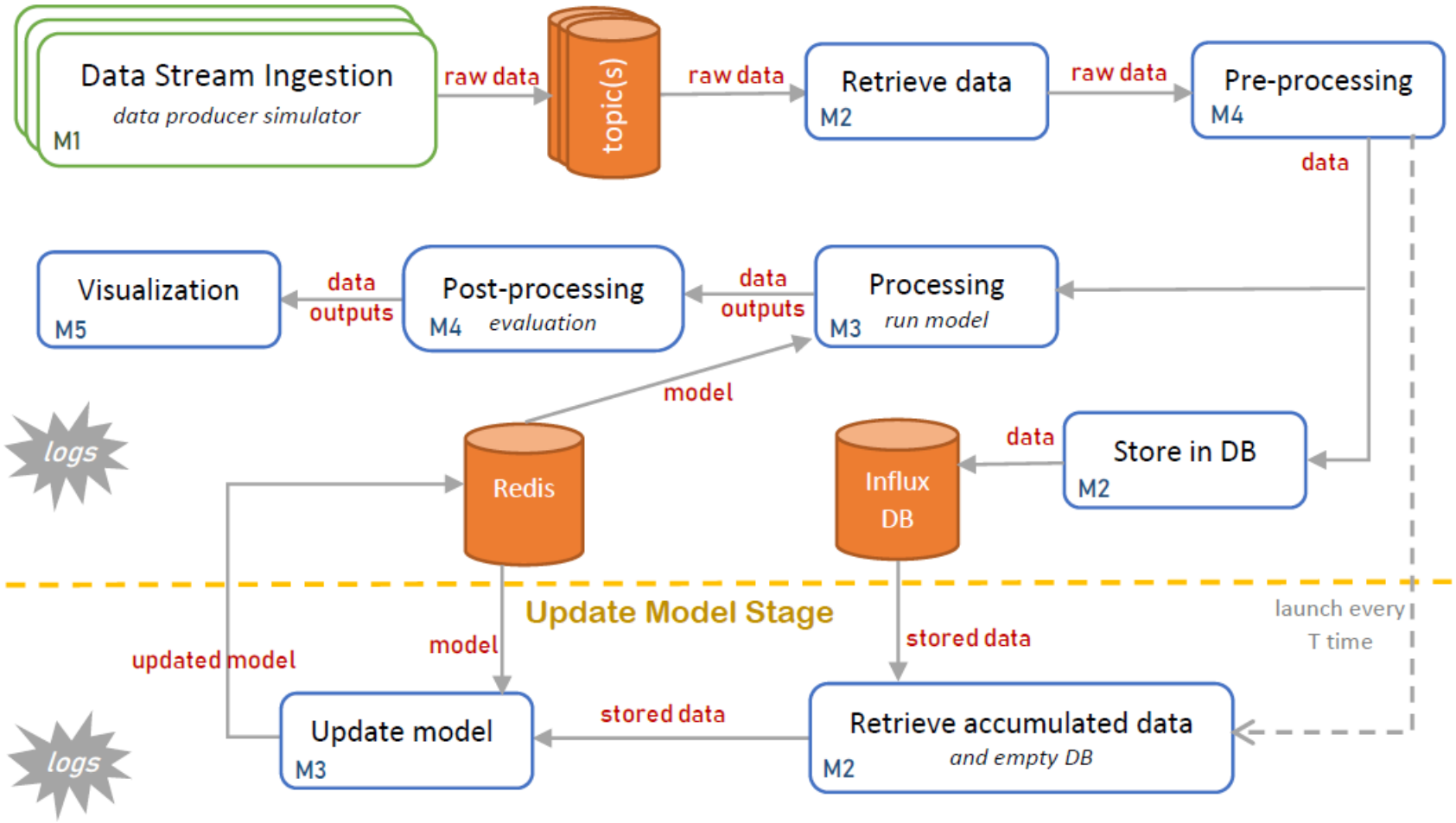


STREAMER ARCHITECTURE

- **STREAMER consists of a layered architecture (independent modules):**
  - M1 – Data Stream Ingestion
    - **Offline** – data are sent by blocks of N records and periodically every t seconds
    - **Online** – data are sent according its timestamps, sending can be scaled, advanced, or delayed in time.
  - M2 – DSPE Processor
    - Drives the streaming pipeline from M1 to M5
  - M3 – Learning API/ API adapter
    - Supports batch training and online training (learning update)
  - M4 – Applications & Evaluation
    - Pre/post processing functionalities & evaluation metrics (accuracy, sensitivity, precision, etc.)
  - M5 – Data Visualization
    - The model's performance can be monitored in a dedicated Kibana dashboard.



STREAMER WORKFLOW



DE LA RECHERCHE À L'INDUSTRIE

cea

www.cea.fr

STREAMER DISSEMINATION

INSTITUT CARNOT CEA LIST

digiteo

list

SCIENTIFIC PUBLICATIONS

- Garcia-Rodriguez, Sandra, Mohammad Alshaer, and Cedric Gouy-Pailler. **"STREAMER: A Powerful Framework for Continuous Learning in Data Streams."** Proceedings of the 29th ACM International Conference on Information & Knowledge Management. 2020. https://dl.acm.org/doi/10.1145/3340531.3417427

- Alshaer, Mohammad, Sandra Garcia-Rodriguez, and Cedric Gouy-Pailler. **"Detecting Anomalies from Streaming Time Series using Matrix Profile and Shapelets Learning."** 2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI). IEEE, 2020. https://www.computer.org/csdl/proceedings-article/ictai/2020/922800a376/1pP3tGGtcn6

- CEA Tech Linkedin = https://www.linkedin.com/feed/update/urn:li:activity:6787752698235224064/
- CEA Tech News (EN) = https://www.cea-tech.fr/cea-tech/english/Pages/2021/streamer-simulates-massive-data-streams.aspx
- CEA Tech News (FR) = https://www.cea-tech.fr/cea-tech/Pages/2021/streamer-simule-des-flux-de-donnees-massives.aspx
- Linkedin (FR/EN) = https://www.linkedin.com/feed/update/urn:li:activity:6787752698235224064/
- Twitter (FR) = https://twitter.com/CEA_List/status/1377182719638003715?s=20
- Annual report (FR) = http://www-list.cea.fr/images/stories/decouvrir-le-list/qui-sommes-nous/rapport-dactivite/Rapport_activite_CEA_LIST_2020.pdf
- CEA LIST News (FR) = http://www-list.cea.fr/medias/toute-l-actualite/2021/496-30-mars-2021-streamer-une-plateforme-logicielle-au-service-de-l-apprentissage-automatique-sur-les-flux-de-donnees
- CEA LIST News (EN) = http://www-list.cea.fr/en/media/news/2021/496-march-30-2021-streamer-a-software-platform-for-machine-learning-for-data-streams
- Paris Saclay University (EN) : https://www.universite-paris-saclay.fr/en/news/streamer-software-platform-machine-learning-data-streams
- DataIA (FR/EN) : https://www.actuia.com/actualite/streamer-un-programme-permettant-dintegrer-et-de-tester-facilement-des-algorithmes-de-machine-learning/

DE LA RECHERCHE À L'INDUSTRIE

cea

www.cea.fr

STREAMER LIFE DEMO

INSTITUT CARNOT CEA LIST

digiteo

list

DEMONSTRATION SCENARIO: ECG

- **Objective**: electrocardiogram (ECG) anomaly detection in time series stream data with no prior labels
  - anomalies may vary in time

- MIT-BIH database
  - 48 ambulatory ECG 30 mins recordings
  - every heartbeat is represented by a different sequence
  - Anomaly: *Ventricular ectopic beat* class

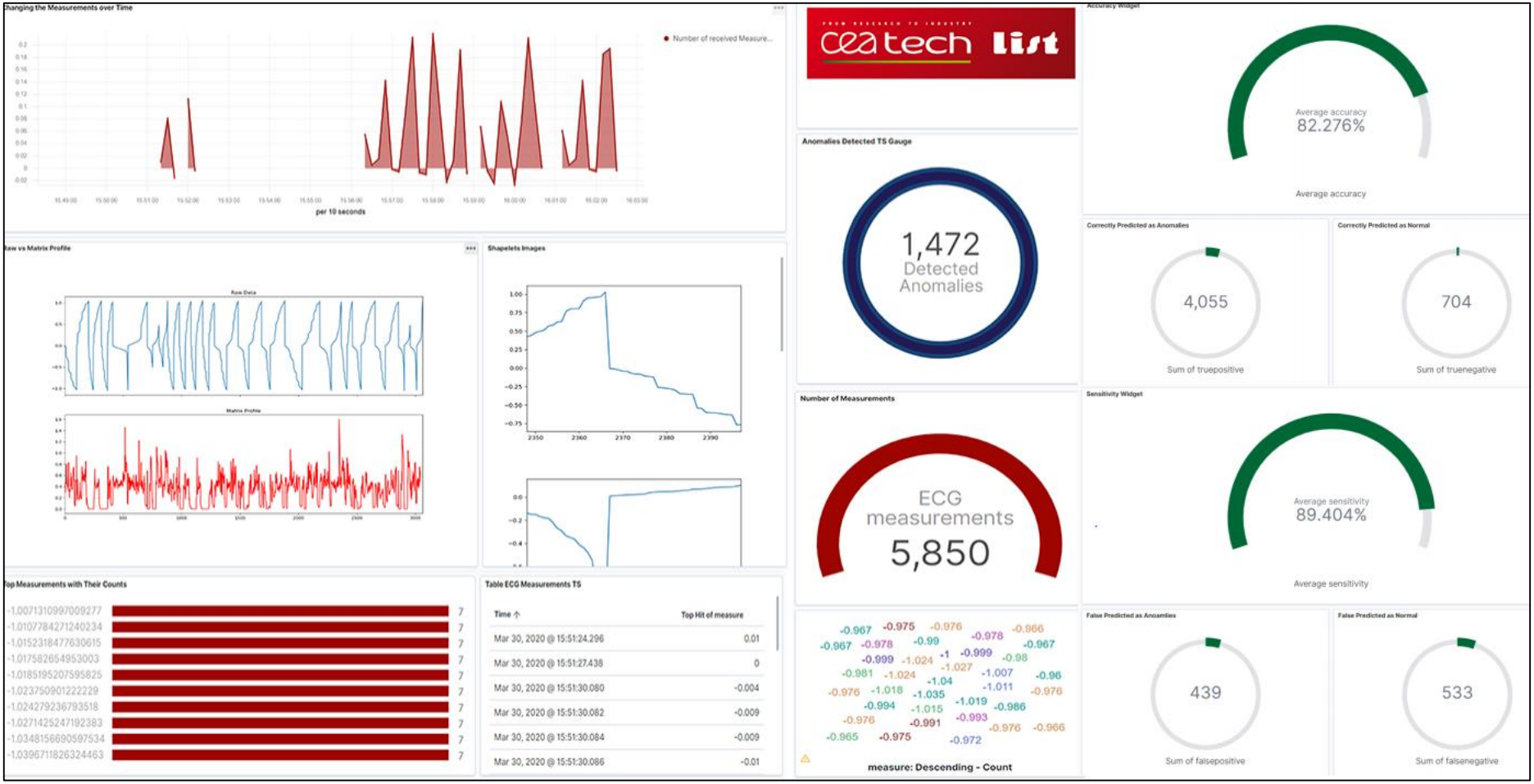
- Unsupervised **learning**
  - *Matrix Profile* algorithm for learning the shapelets (features) that represent anomalies over unsupervised data streams
  - Based on them, we do the detection of the anomalies by performing a similarity function
  - Model is updated after every bunch of data received

- **Evaluation**: labels of MIT-BIH are only used for metrics calculation

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Sensitivity = \frac{TP}{TP + FN}$$

where TP means true positive, TN true negative, FP false positive and FN false negative.



Demonstration video is found at the following link:
https://www.youtube.com/watch?v=UHx1YenTF1Y&feature=emb_title

DEMONSTRATION SCENARIO: 2 USE CASES

- **STREAMER** can run several scenarios in parallel
  - Different configuration files for each UC

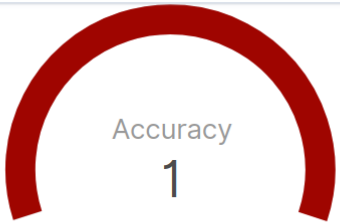
- Network intrusion detection (KDD)
  - Classification (23 classes)
  - Algorithm: Hoeffding Tree Classifier
  - Metrics: accuracy, F1 score, precision & sensitivity

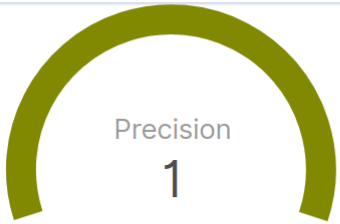
- Demand forecasting (Multivariate time series)
  - Regression
  - Algorithm: XGBoost
  - Metrics: MSE, RMSE, MAE, R2

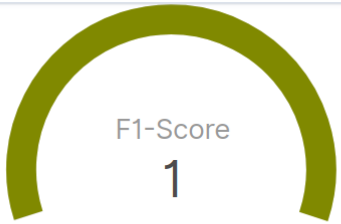
DEMONSTRATION SCENARIO: 2 USE CASES

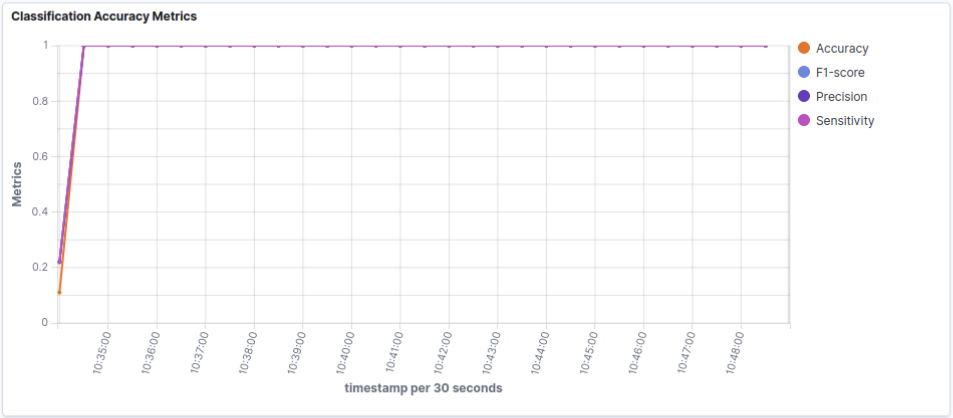
Dashboard / Performance Evaluation Dashboard

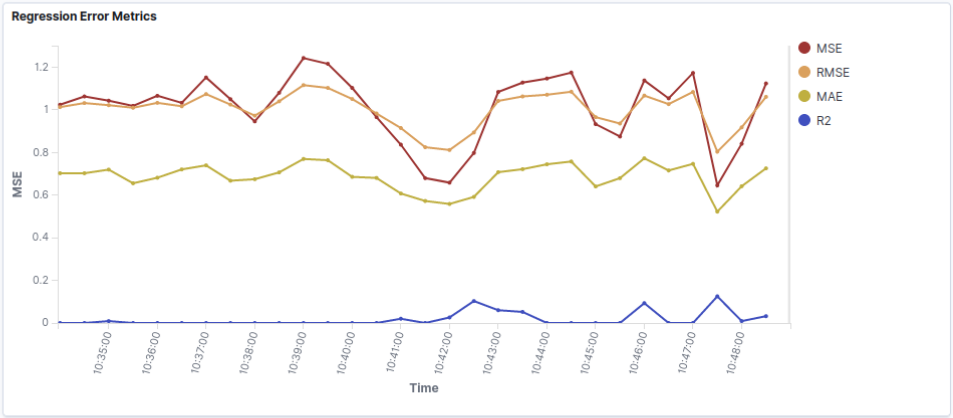
Full screen | Share | Clone | Edit

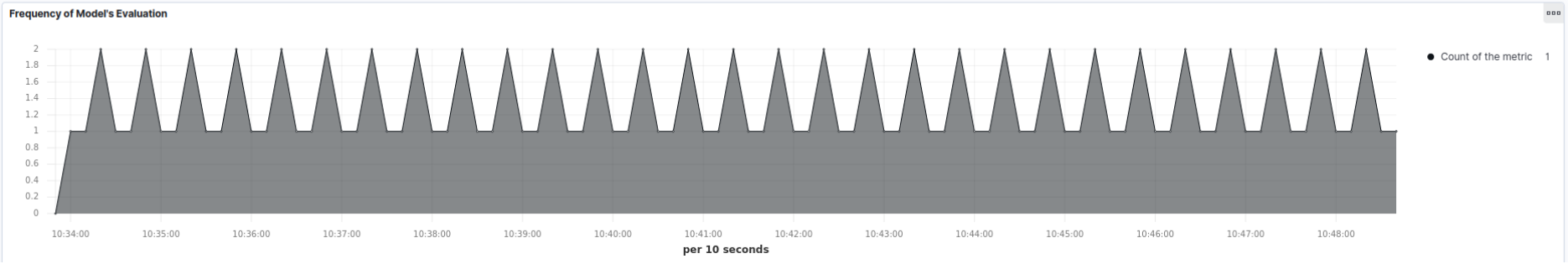












STREAMER

a Powerful and Open-Source Framework for Continuous Learning in Data Streams

Commissariat à l'énergie atomique et aux énergies alternatives
Institut Carnot CEA LIST
CEA Saclay – DIGITEO LABS | 91191 Gif-sur-Yvette Cedex
Phone. +33 (0)1 69 08 88 55 | Fax. +33 (0)1 69 08 89 65

Direction: CEA TECH
Department: DCSI
Laboratory: LI3A

Etablissement public à caractère industriel et commercial | RCS Paris B 775 685 019