Thesis Proposals

Streaming Data Analytics

Prof. Emanuele Della Valle

December 2022

Team

Emanuele Della Valle

Associate Professor

Stream Reasoning, Machine Learning, Time Series Analysis & IoT

Alessio Bernardo

PhD Candidate

SML for imbalanced and evolving streams

Giacomo Ziffer

PhD Student

SML for evolving time-dependent data streams

Federico Giannini

PhD Student

Neural networks for evolving time-dependent data streams

Type of theses

Streaming Data Engineering

Streaming Data Integration

- Industrial theses
- Research theses

Stream Graph Processing

Streaming Data Science

Streaming Machine Learning

- Industrial theses
- Research theses

Streaming Data Integration

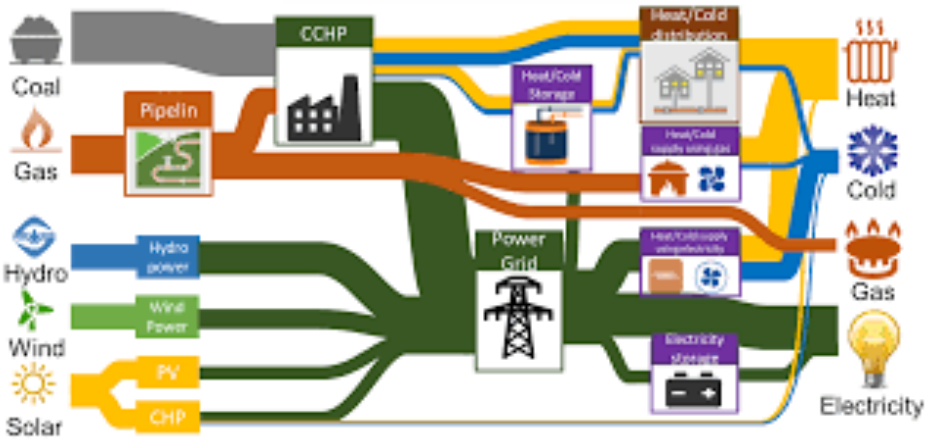
Industrial theses

RSE – Ricerca Sistema Energetico

Multi Energy Streaming Data Analytics

Goal

Empirically demonstrate that our Chimera Suite (https://github.com/chimera-suite/) helps in performing descriptive and predictive analysis of heterogenous data streams captured monitoring a multi energy power supply networks (electricity, gas, steam, …) exploiting a multidomain industrial knowledge graph



Challenging due to:

- The need to enable streaming data analytics at scale while reducing the effort dedicated to building data pipeline and increase the coherence bet

Streaming Data Integration

Research theses

Model the execution semantics of Stream Processing Engines to enable their integration

Goal

Given the same continuous query registered on the same data streams, existing Stream Processing Engines (e.g., Spark, Flink, ksqldb) give different answers. Integrating them to build applications that use existing continuous data pipelines is challenging. The thesis aims at modelling the execution semantics of those engines and using the model to explain the differences and automatically reconcile them.

Challenging due to:

- identify minimal examples that showcase the difference between the engines
- define a model that balances comprehensiveness and usefulness
- show that such a model enables the integration of existing continuous data pipelines without errors

OntopStream: a Streaming Virtual Knowledge Graph

Goal

Ontop (https://ontop-vkg.org/) is a popular Virtual Knowledge Graph System. In a previous work, we laid the foundation of its streaming extension with OntopStream (https://github.com/chimera-suite/OntopStream). The current prototype supports the continuous execution of a single query against a streaming virtual knowledge graph used to integrate multiple heterogeneous streams. In this thesis, we want to further extend out previous work making it possible to register multiple queries with multiple windows.

Challenging due to:

- understand virtual knowledge graph technology
- familiarise with the internals of Ontop to contribute to an open
- empirically demonstrate that the system works and scales as expected

Streaming Machine Learning

Industrial theses

RSE – Ricerca Sistema Energetico

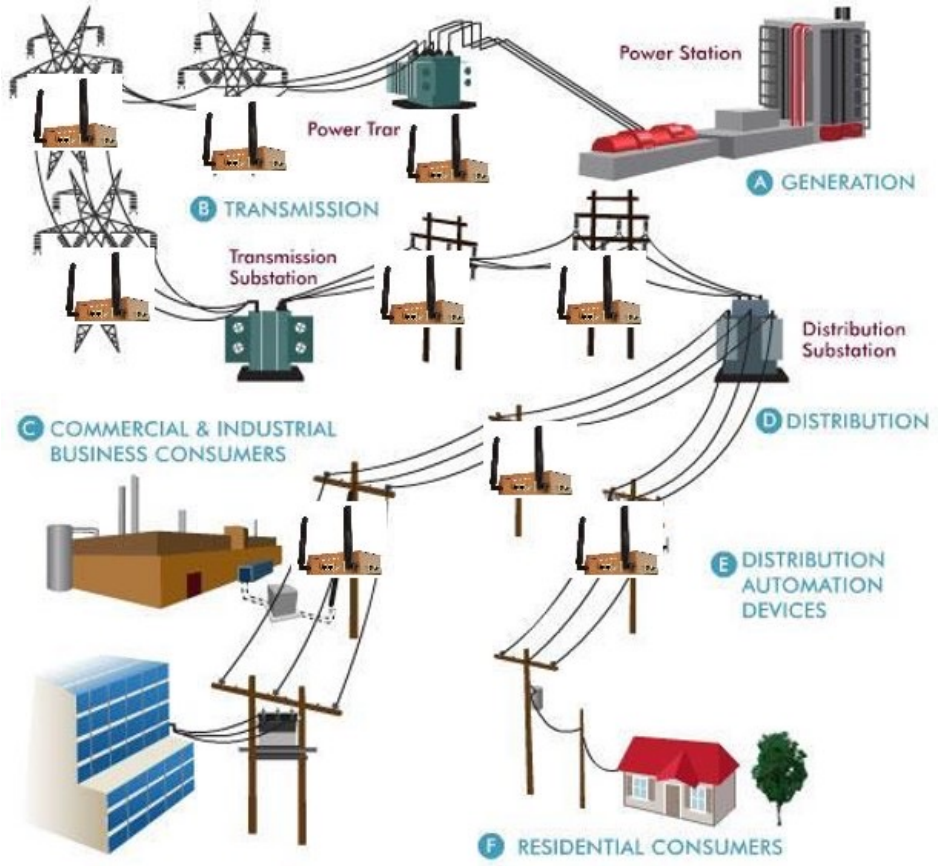
Fault detection in electrical network

Goal

Predict any fault in the electrical distribution network based on the transformers characteristics (power, reactive power, voltage) and weather conditions (temperature, rain, wind)

Challenging due to:

- the high reactiveness that the system must have: there must be enough time between when the failure is predicted and when it really happens to allow the technicians to avoid the problem



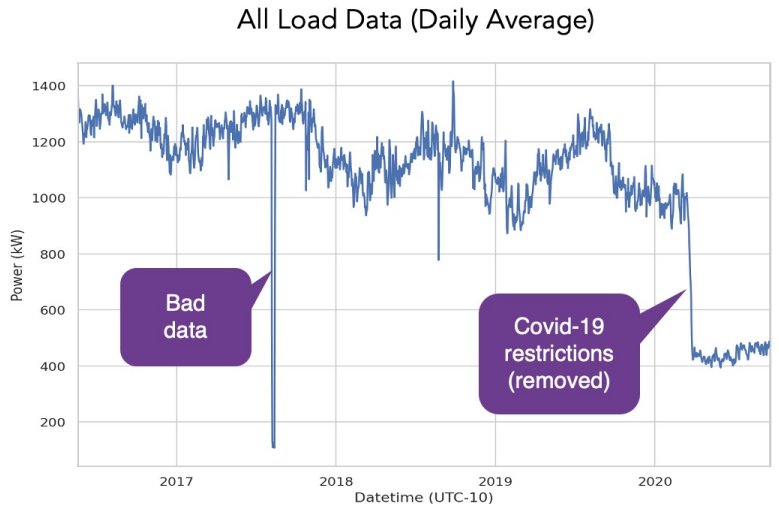
Microgrid electric load forecasting

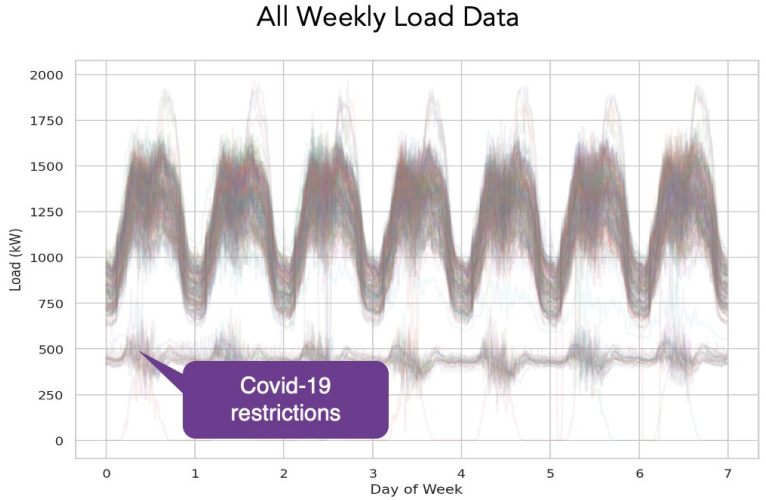
Goal

Forecast the day-ahead electric load of a hotel to optimize building microgrid activities

Challenging due to:

- lack of spatial aggregation (compared to sub-station scale) results in a highly non-linear and variable time series
- can't assume exogenous attributes available one day ahead
- changes caused by Covid-19 restrictions





Von Krannichfeldt, L., Wang, Y., & Hug, G. (2020). **Online ensemble learning for load forecasting**. IEEE Transactions on Power Systems, 36(1), 545-548.

Streaming Machine Learning

Research theses

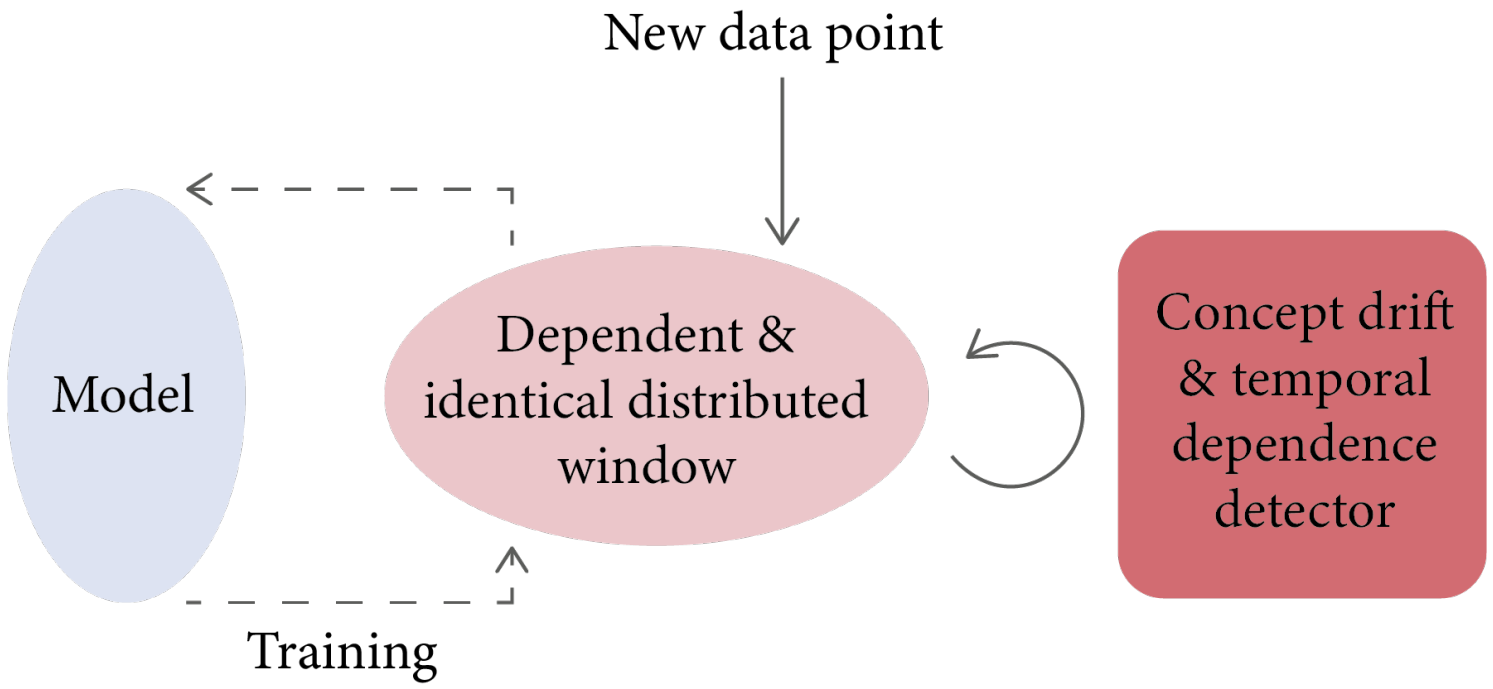
Concept drift detector for not i.i.d. data streams

Setting

In many data streams, temporal dependence is part of the concept and data are not i.i.d. This can hugely affect the model performance.

Goal

How to effectively exploit this peculiar problem together with the concept drift detection?



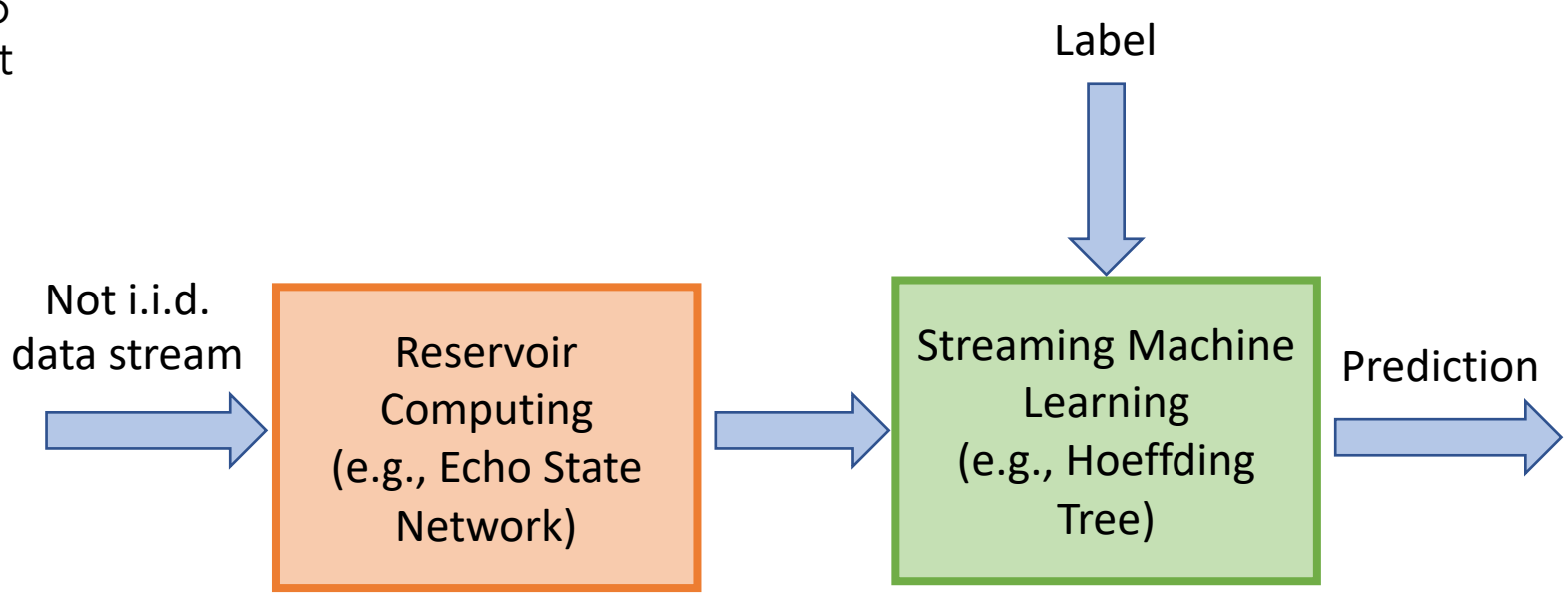
Streaming Reservoir Computing

Setting

SML algorithms are powerful and easy to deploy (no hyper-parameter to tune), but **unable to capture strong temporal dependences**. Reservoir Computing offers algorithms derived from RNN that require few computations and capture time dependence.

Goal

How can we combine Reservoir Computing and SML to learn temporal dependences in data streams in real-time?



SML for Lifelong Learning

Setting

SML algorithms learn online and can adapt to changes, however they suffer from catastrophic forgetting in class-incremental lifelong learning scenarios.

Goal

How to combine adaptation to new data and retention of useful past information?

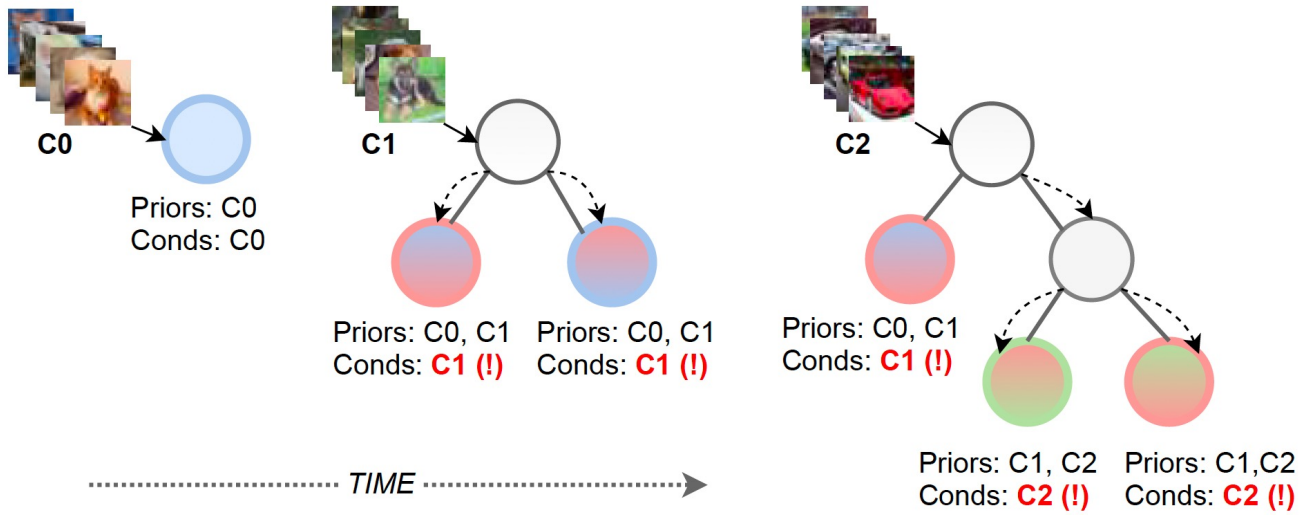


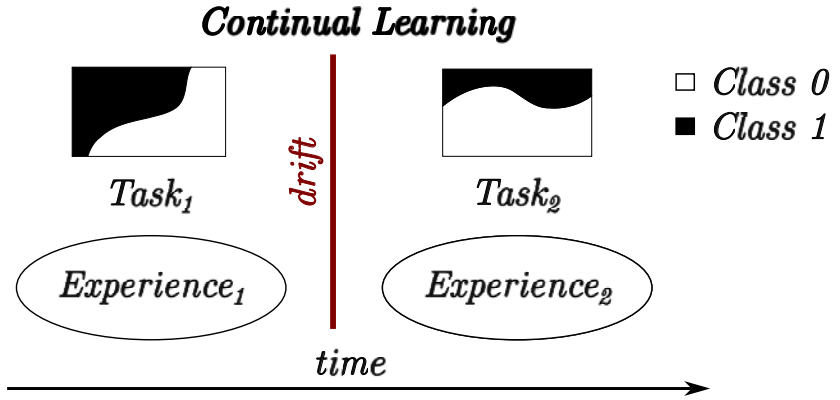
Fig. 1: Catastrophic forgetting in streaming decision trees learning from a class-incremental sequence.

Łukasz Korycki and Bartosz Krawczyk. 2021. **Streaming Decision Trees for Lifelong Learning**. In Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part I. Springer-Verlag, Berlin, Heidelberg, 502–518.

CL Architectural strategies for SML

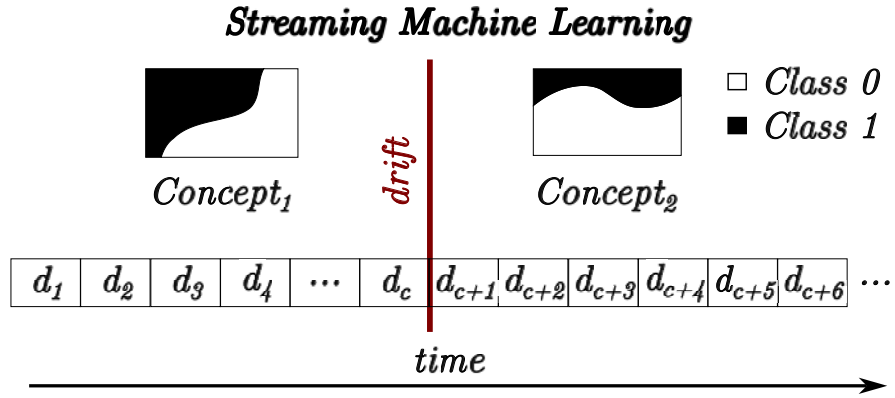
Setting

Architectural strategies are applied in CL scenarios to avoid catastrophic forgetting and use transfer learning to adapt quickly to new tasks. They require **task labels** associated with the data points to be used. In the SML context, there are many methodologies to detect drifts in a stream, but Neural Networks are less studied.



Goal

How can we efficiently mix CL architectural strategies and concept drift detectors to apply Neural Networks in an SML context and avoid catastrophic forgetting?



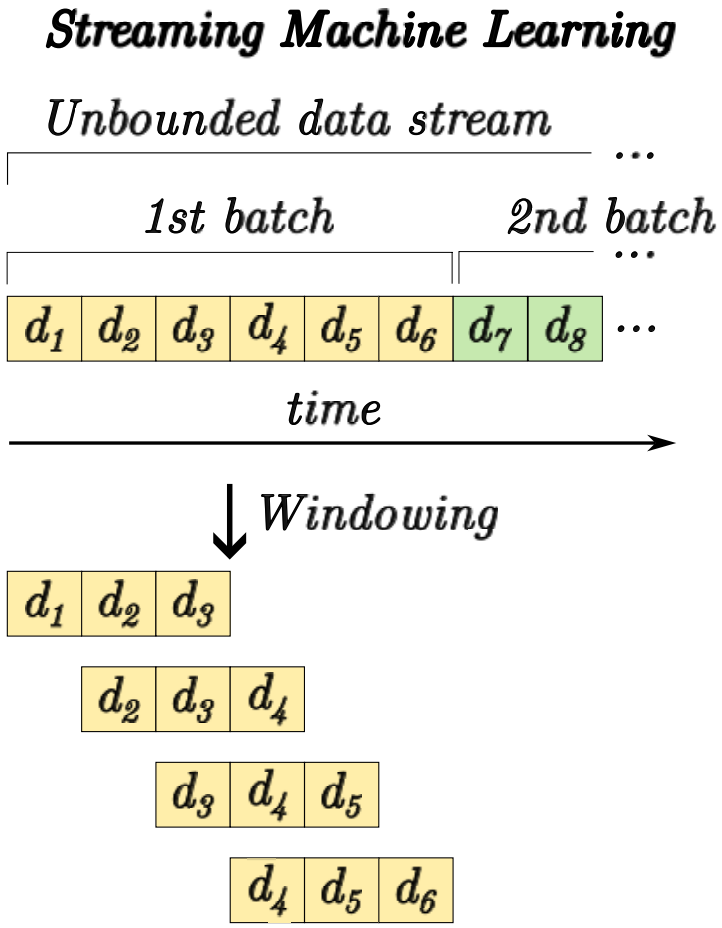
CL Architectural strategies for Evolving Streaming Time Series

Setting

In many data streams, temporal dependencies are part of the concept and data are not i.i.d. This can hugely affect the model performance. Recurrent Neural Networks (RNN) are Neural Networks meant to tame temporal dependencies.

Goal

How can we adapt Architectural strategies to RNNs and use them with concept drift detectors in order to jointly manage concept drifts, catastrophic forgetting and temporal dependencies?



Internship

Internship industrial projects

BOSCH

RSE
Ricerca
Sistema
Energetico

Rai Way

EURA NOVA

NTT DATA

SIEMENS

THALES

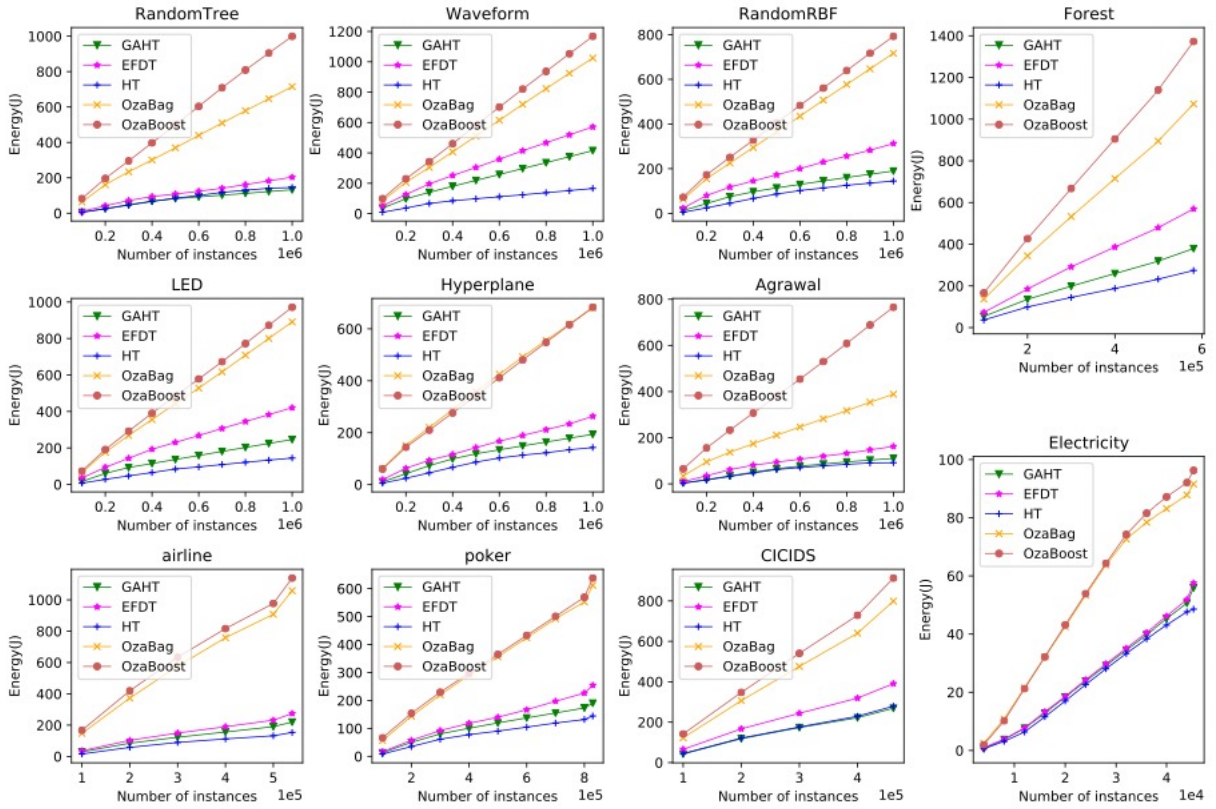
ThalesAlenia
a Thales / Leonardo company
Space

Siemens
TinyOL

SIEMENS

Goal

Further develop TinyOL, a library based on TinyML that enables incremental training on streaming data directly on MCU



Ren, H., Anicic, D., & Runkler, T. A. (2021, July). **Tinyol: Tinyml with online-learning on microcontrollers.** In 2021 International Joint Conference on Neural Networks (IJCNN) (pp. 1-8). IEEE.

Other information

- Thesis time span: 8 – 14 months
- Tesina time span: 4 – 6 months
- If you are interested in one of these topics, fill this form:
  - https://forms.office.com/e/kbNndxLBC0
- Deadline: 01/09/2023

Thesis Proposals

Streaming Data Analytics

Prof. Emanuele Della Valle

December 2022