Streaming Data Analytics
Thesis proposal

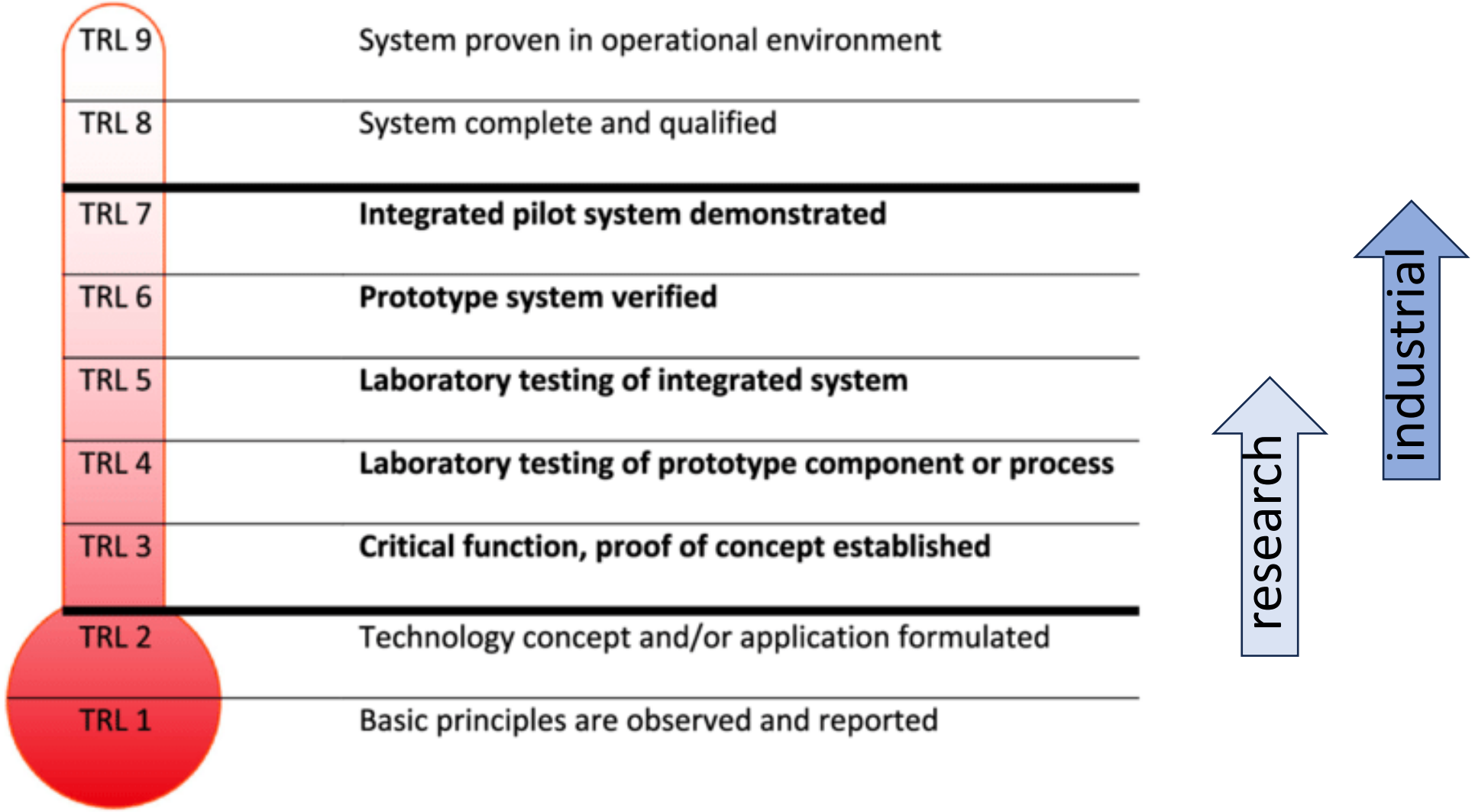
Emanuele Della Valle
Politecnico di Milano

POLITECNICO
MILANO 1863

Agenda

- SDA theses
- Joint theses with foreign university
- Internships

Technology Readiness Levels



State of the Art vs. State of the Practice

State of the Art

- the **highest level of development of a** technique or **scientific field**, achieved at a particular time
- In our context, it refers to the latest and most advanced streaming data engineering and streaming data science **academic results**

State of the Practice

- the current accepted methodologies, techniques, tools, and approaches that are **widely used in a particular industry**
- In our context, it involves using widely accepted systems, tools, and methodologies that are **commonly used by Data Engineers and Data Scientists** in their day-to-day processing of data streams

Theses's types

	Novel result	Benchmarking	Use case
Industrial	Develop a novel algorithm/component so that the result can be the core of a **new product/service** **TRL 5**	**Compare** multiple **state-of-the-practice solutions** in several real-world conditions to determine the **trade-offs** **TRL 6**	Investigate which is the best way to **address a real-world problem** using **state-of-the-practice** solutions **TRL 7**
Research	Develop a novel algorithm/method so that the result can be the content of a **scientific publication** **TRL 3**	**Compare** multiple **state-of-the-art methods** in several previously published settings to determine the **trade-offs** **TRL 4**	Investigate which is the best way to address a real-world problem using **state-of-the-art** solutions **TRL 5**

Theses's types

	Novel result	Benchmarking	Use case
Industrial	Develop a novel algorithm/component so that the result can be the core of a **new product/service** **TRL 5**	**Compare** multiple **state-of-the-practice solutions** in several real-world conditions to determine the **trade-offs** **TRL 6**	Investigate which is the best way to **address a real-world problem** using **state-of-the-practice** solutions **TRL 7**
Research	Develop a novel algorithm/method so that the result can be the content of a **scientific publication** **TRL 3**	**Compare** multiple **state-of-the-art methods** in several previously published settings to determine the **trade-offs** **TRL 4**	Investigate which is the best way to address a real-world problem using **state-of-the-art** solutions **TRL 5**

Theses's types

	Novel result	Benchmarking	Use case
Industrial	Develop a novel algorithm/component so that the result can be the core of a **new product/service** **TRL 5**	**Compare** multiple **state-of-the-practice solutions** in several real-world conditions to determine the **trade-offs** **TRL 6**	Investigate which is the best way to **address a real-world problem** using **state-of-the-practice** solutions **TRL 7**
Research	Develop a novel algorithm/method so that the result can be the content of a **scientific publication** **TRL 3**	**Compare** multiple **state-of-the-art methods** in several previously published settings to determine the **trade-offs** **TRL 4**	Investigate which is the best way to address a real-world problem using **state-of-the-art** solutions **TRL 5**

Theses's types

	Novel result	Benchmarking	Use case
Industrial	Develop a novel algorithm/component so that the result can be the core of a **new product/service** **TRL 5**	**Compare** multiple **state-of-the-practice solutions** in several real-world conditions to determine the **trade-offs** **TRL 6**	Investigate which is the best way to **address a real-world problem** using **state-of-the-practice** solutions **TRL 7**
Research	Develop a novel algorithm/method so that the result can be the content of a **scientific publication** **TRL 3**	**Compare** multiple **state-of-the-art methods** in several previously published settings to determine the **trade-offs** **TRL 4**	Investigate which is the best way to address a real-world problem using **state-of-the-art** solutions **TRL 5**

Theses's types

	Novel result	Benchmarking	Use case
Industrial	Develop a novel algorithm/component so that the result can be the core of a **new product/service** **TRL 5**	**Compare** multiple **state-of-the-practice solutions** in several real-world conditions to determine the **trade-offs** **TRL 6**	Investigate which is the best way to **address a real-world problem** using **state-of-the-practice** solutions **TRL 7**
Research	Develop a novel algorithm/method so that the result can be the content of a **scientific publication** **TRL 3**	**Compare** multiple **state-of-the-art methods** in several previously published settings to determine the **trade-offs** **TRL 4**	Investigate which is the best way to address a real-world problem using **state-of-the-art** solutions **TRL 5**

Theses's types

	Novel result	Benchmarking	Use case
Industrial	Develop a novel algorithm/component so that the result can be the core of a **new product/service** **TRL 5**	**Compare** multiple **state-of-the-practice solutions** in several real-world conditions to determine the **trade-offs** **TRL 6**	Investigate which is the best way to **address a real-world problem** using **state-of-the-practice** solutions **TRL 7**
Research	Develop a novel algorithm/method so that the result can be the content of a **scientific publication** **TRL 3**	**Compare** multiple **state-of-the-art methods** in several previously published settings to determine the **trade-offs** **TRL 4**	Investigate which is the best way to address a real-world problem using state-of-the-art solutions TRL 5

Theses's types

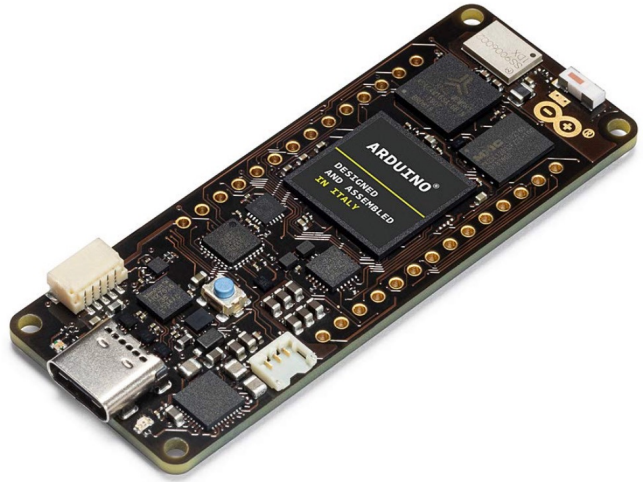
	Novel result	Benchmarking	Use case
Industrial	Develop a novel algorithm/component so that the result can be the core of a **new product/service** **TRL 5**	**Compare** multiple **state-of-the-practice solutions** in several real-world conditions to determine the **trade-offs** **TRL 6**	Investigate which is the best way to **address a real-world problem** using **state-of-the-practice** solutions **TRL 7**
Research	Develop a novel algorithm/method so that the result can be the content of a **scientific publication** **TRL 3**	**Compare** multiple **state-of-the-art methods** in several previously published settings to determine the **trade-offs** **TRL 4**	Investigate which is the best way to address a real-world problem using **state-of-the-art** solutions **TRL 5**

Extending and Optimizing streamDM-C++ Library for Efficient Streaming ML Algorithms on Microcontrollers

Industrial novel result thesis

Goal

In an Edge Computing context, expand StreamDM-C++ library with further SML algorithms, optimize it to be executed on a MCU (Arduino Portenta H7), and test the performance w.r.t. the MOA and River libraries.



motus ml

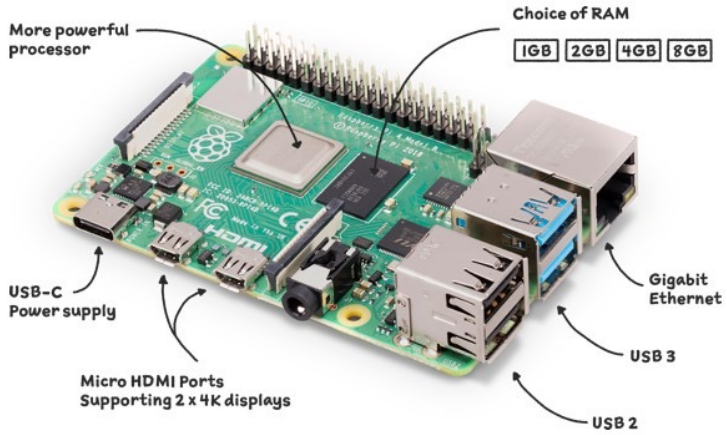
spin off® POLITECNICO DI MILANO

Suitability analysis of the Raspberry Pi to Run Streaming Machine Learning Algorithms

Industrial novel result thesis

Goal

In an Edge Computing context, optimize the MOA library, through quantization, to run several tests with different streams and algorithms on a Raspberry Pi 4b.



motus ml

spin off® POLITECNICO DI MILANO

Performance analysis of different implementations of the Kappa Architecture

Industrial benchmarking thesis

Goal

The real-world performance of the Kappa architecture can vary significantly depending on its implementation and deployment.

The goal of the thesis is to provide a comprehensive evaluation of the Kappa architecture's ability to handle different kinds of streaming data sources in real-world scenarios, addressing the lack of a thorough analysis that evaluates the end-to-end performance of a streaming processing pipeline.

CDC technology	Max throughput [records/s]
OSS Debezium	57395
Managed Debezium (Confluent Cloud)	12764
DMS (AWS)	56652

	Matched ride event		
Latency [ms]	Average	Max	Min
Kafka	421.137	929	122
Redpanda	773.222	1934	401
Pulsar	1346.670	2166	981
Confluent	644.441	1098	355
AWS	1293.593	2118.561	831.559

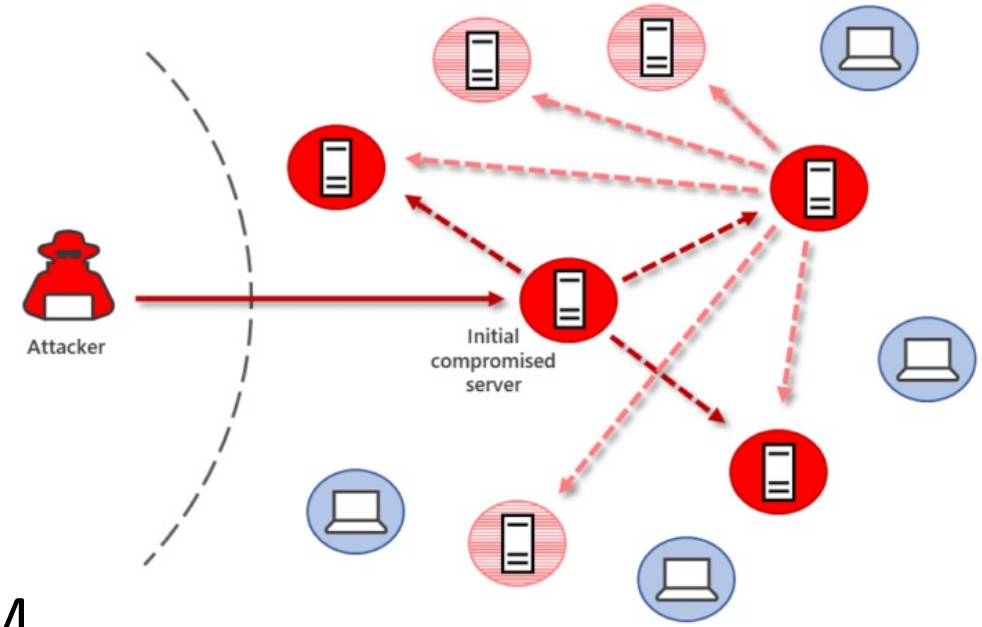
Towards an Unsupervised graph-based Machine Learning model for Lateral Movement Detection

Industrial/Academic thesis on a specific use case

Goal

Reimplement unsupervised graph-based Machine Learning mode to

- confirm its ability to detect anomalous edges appearing over a streaming graph
- investigate its applicability to detect lateral movements (a type of cyberattack)
- optimize its performances to allow its applicability to a real-world industrial setting (1M events per day)

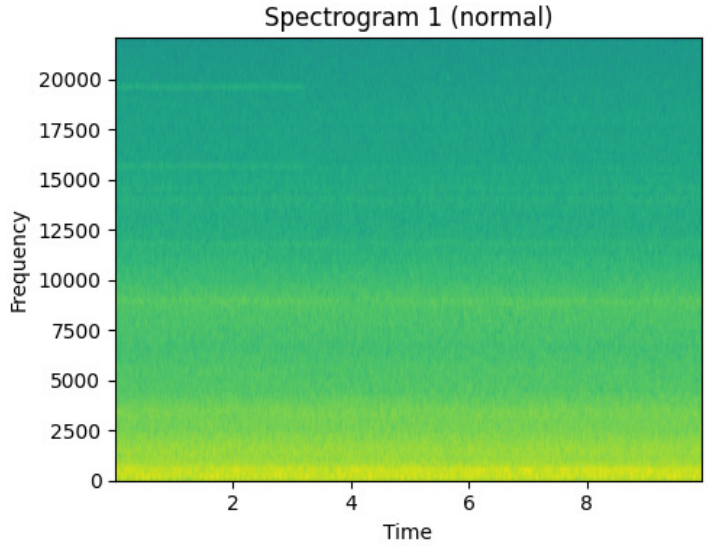


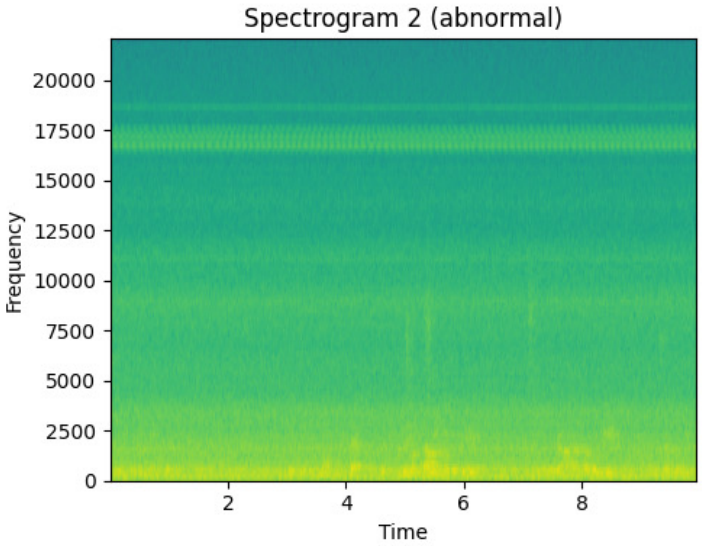
Streaming Anomaly Detection for Cooling Pumps' noise

Industrial thesis on a specific use case

Goal

Test several configurations of streaming autoencoder methods to identify anomalous sounds of cooling pumps.

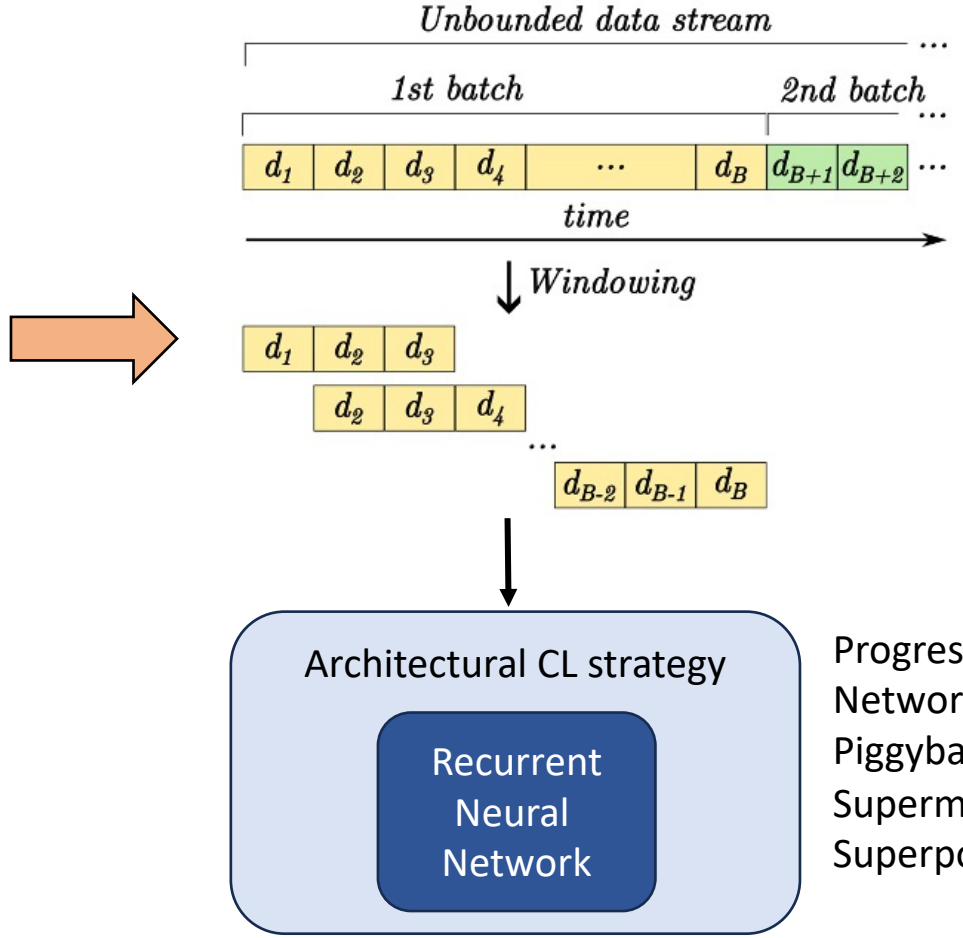




Blending Continual Learning, Streaming Machine Learning and Time Series Analysis

Research novel result thesis

	SML	CL	TSA
Learning continuously	X		
Managing **concept drifts**.	X	X	
Taming **temporal dependence**.			X
Avoiding **catastrophic forgetting**.		X	



Progressive Neural Networks[1]
Piggyback
Supermarks in Superposition

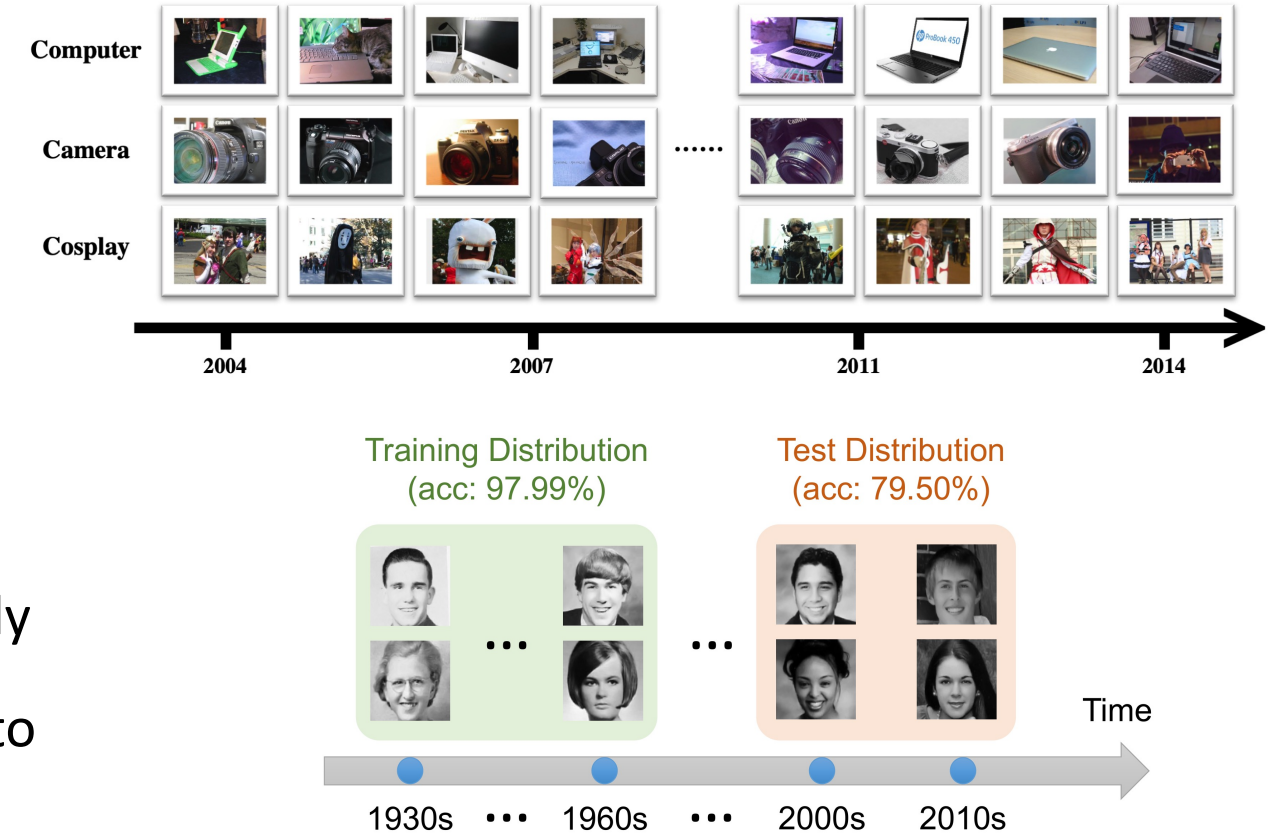
1) Giannini, F., Ziffer, G., & Della Valle, E. (2023). cPNN: **Continuous Progressive Neural Networks for Evolving Streaming Time Series**. In Pacific-Asia Conference on Knowledge Discovery and Data Mining (pp. 328-340).

Streaming Machine Learning for Classification of Temporally-Evolving Images

Research novel result thesis

Goal

Develop SML classification algorithms capable of addressing real temporal distribution shifts in datasets. The primary goal is to enable SML models that learn from real-world images exhibiting **smooth temporal evolution.**
The developed approaches must continually learn from **not i.i.d. unstructured data**, while handling concept drift and adapting to the **temporal coherence** of the stream.

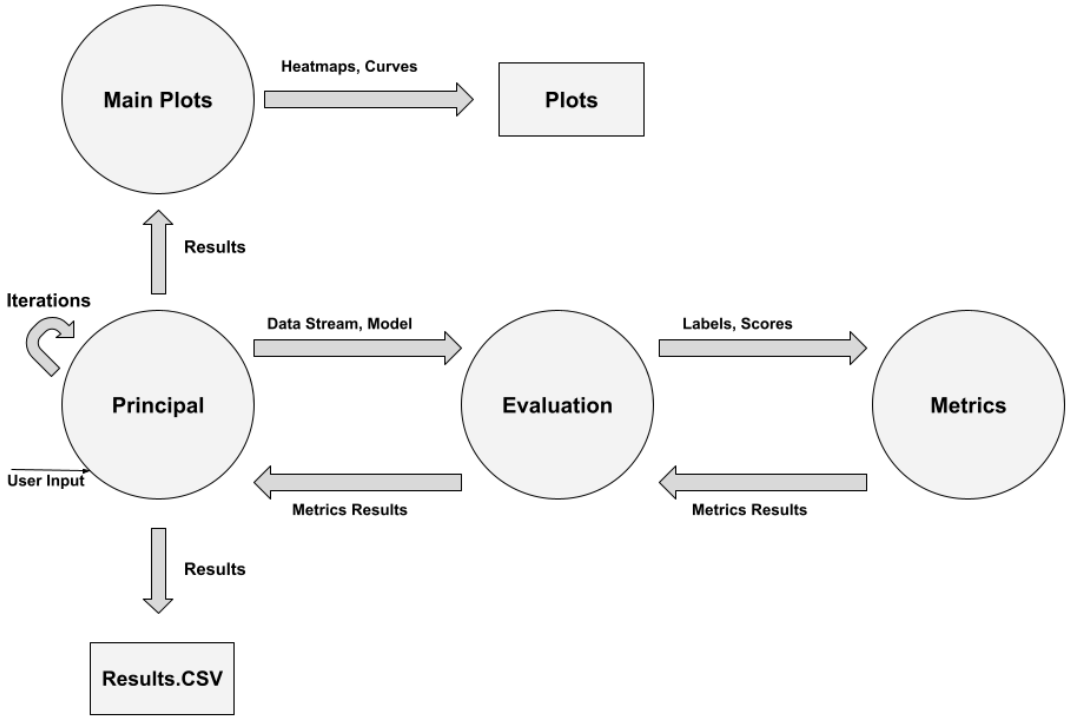


ADBench: a Novel Benchmark for Streaming Anomaly Detection Algorithms

Research benchmarking thesis

Goal

Create a benchmarking environment to evaluate the performance of state-of-the-art anomaly detection algorithms in streaming data scenarios, comparing models across various streams, evaluation metrics, visualization techniques, and statistical testing.



Comparing Traditional and Streaming Machine Learning Methods for Soccer Pass Detection

Research thesis on a specific use case

Goal

Compare the ability of traditional ML and SML models in distinguishing pass from non-pass actions using only the footballers' leg movements in the case of balanced, progressively imbalanced, and rebalanced datasets.

Joint theses with foreign university

Implementing Edge's Properties in Quine

Industrial novel result thesis

Goal

Implement edge properties in the context of Quine, an open-source streaming graph engine data that partially supports the Labelled Property Graph data model. The thesis is sponsored by thatDot via Insa valor. Quine is written in Scala and uses AKKA

INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES LYON

thatDot

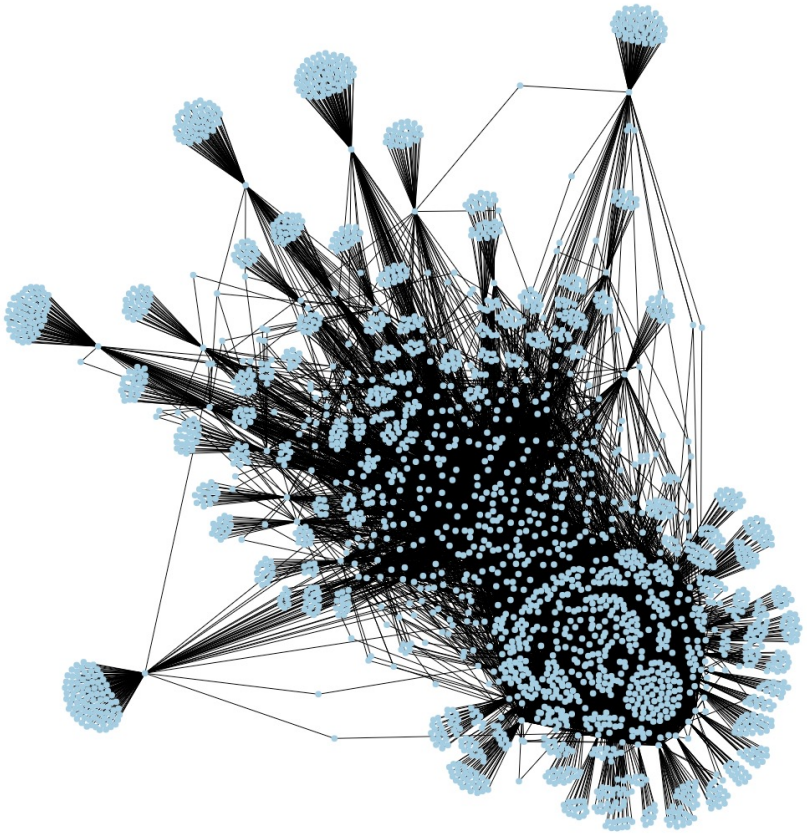
http://thatdot.com/

Study of the Temporal Evolution of IMKG

Research novel result thesis

Goal

Study the evolutionary patterns of entity, usage, and evolution of Internet Memes using our dedicated Knowledge Graph. A descriptive analysis will be the result of our investigation.



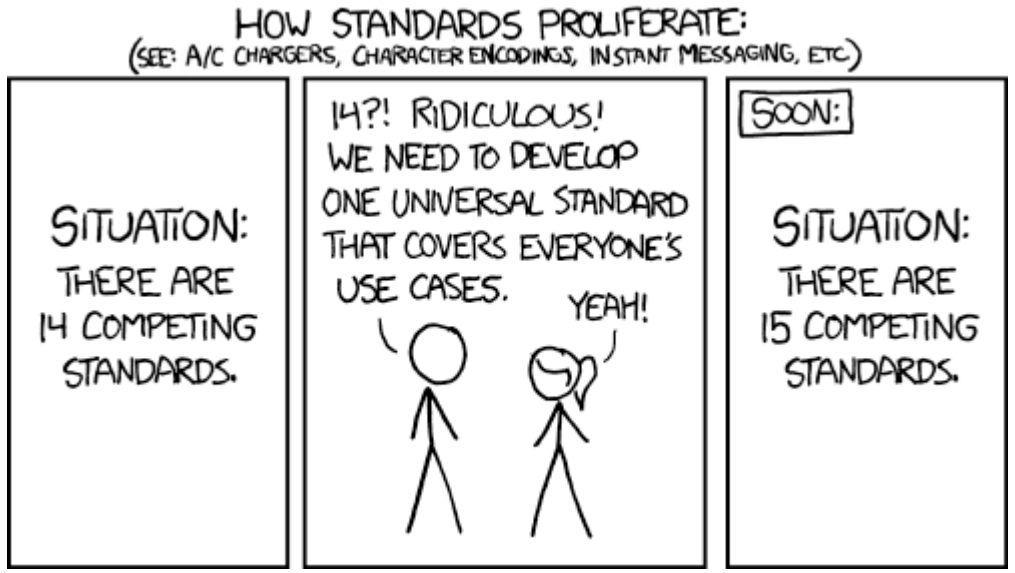
https://memes.science/

Systematic and Comparative Analysis of Declarative Streaming Languages

Research benchmarking thesis

Goal

Catalog, study, and compare the numerous languages for Stream Processing (.e.g, Streaming SQL) and organize them into a comprensive taxonomy. The work can extend to theoretical foundation of continuous queries.



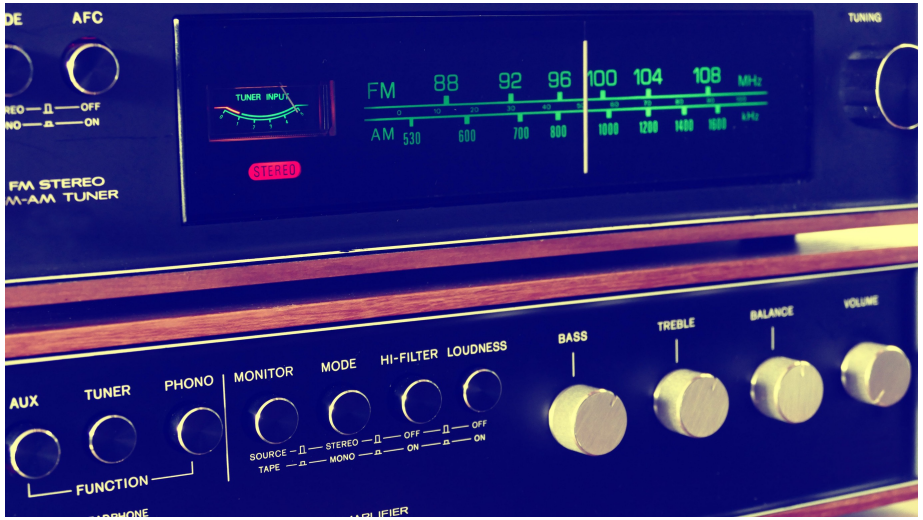
http://streaminglangs.io/

Automated Machine Learning for Data Streams

Research novel result thesis

Goal

Propose a new framework for finding well-performing SML models and their corresponding configurations without the need of machine learning experts.



Internships

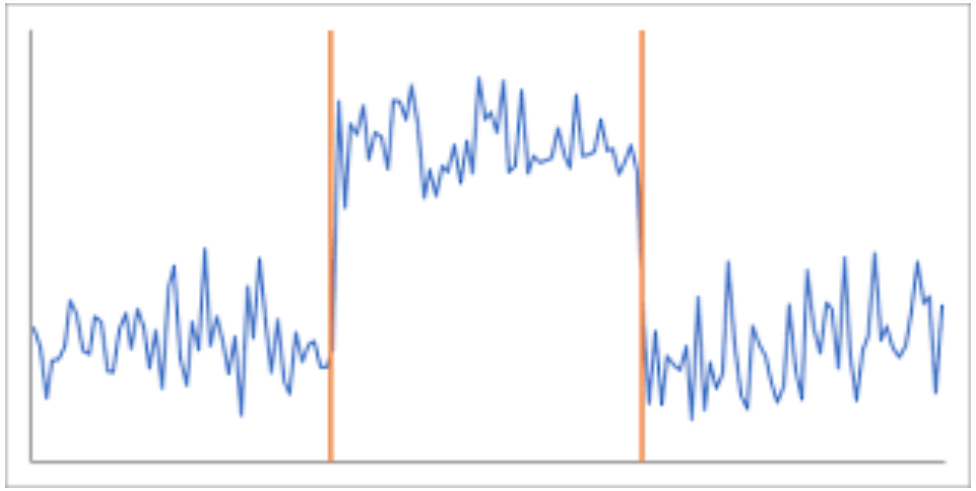
Early Detection of Concept Drift

EURA NOVA research center

Industrial novel result thesis

Goal

Explore the literature on concept drift detection in time series data and propose a new method for triggering alarms when new measurements are gathered over time.



You can find more information in Euranova website*

*The internship can be transformed in a thesis

Other information

If you are interested in one of these topics,

- Fill this form: https://forms.office.com/e/SqajbZNqgg
- Or scan the Qrcode:

Streaming Data Analytics
Thesis proposal

Emanuele Della Valle
Politecnico di Milano

POLITECNICO
MILANO 1863