Streaming Machine Learning Introduction

Alessio Bernardo

Post-doc @ Politecnico di Milano

CTO & Co-founder @ Motus ml

POLITECNICO
MILANO 1863

Me

Alessio Bernardo, Ph.D.

Post-doc @ Politecnico di Milano
CTO & Co-founder @ Motus ml

- **5 years** of experience in **research** in the **Streaming Machine Learning** field with **evolving** data **streams**, **concept drifts**, and **class imbalance;**
- Focus on applying **Streaming Machine Learning** techniques in **constrained** environments at the network's **edge.**

https://alessiobernardo.github.io/
https://motusml.com/

It is a Streaming World …

It is a Streaming World …

Off-shore oil operations

Smart Cities

Power turbine

Social networks

Generate data streams!

E. Della Valle, S. Ceri, F. van Harmelen, D. Fensel **It's a Streaming World! Reasoning upon Rapidly Changing Information**. IEEE Intelligent Systems 2009

… looking for reactive answers …

When a sensor on a drill indicates that it is about to get stuck, how long can I keep drilling?

Where am I likely going to run into a traffic jam during my commute tonight?

Which electrical turbine has sensor readings like any turbine that had a critical failure?

Who is driving the discussion about the top 10 emerging topics ?

Require continuous processing and reactive answer

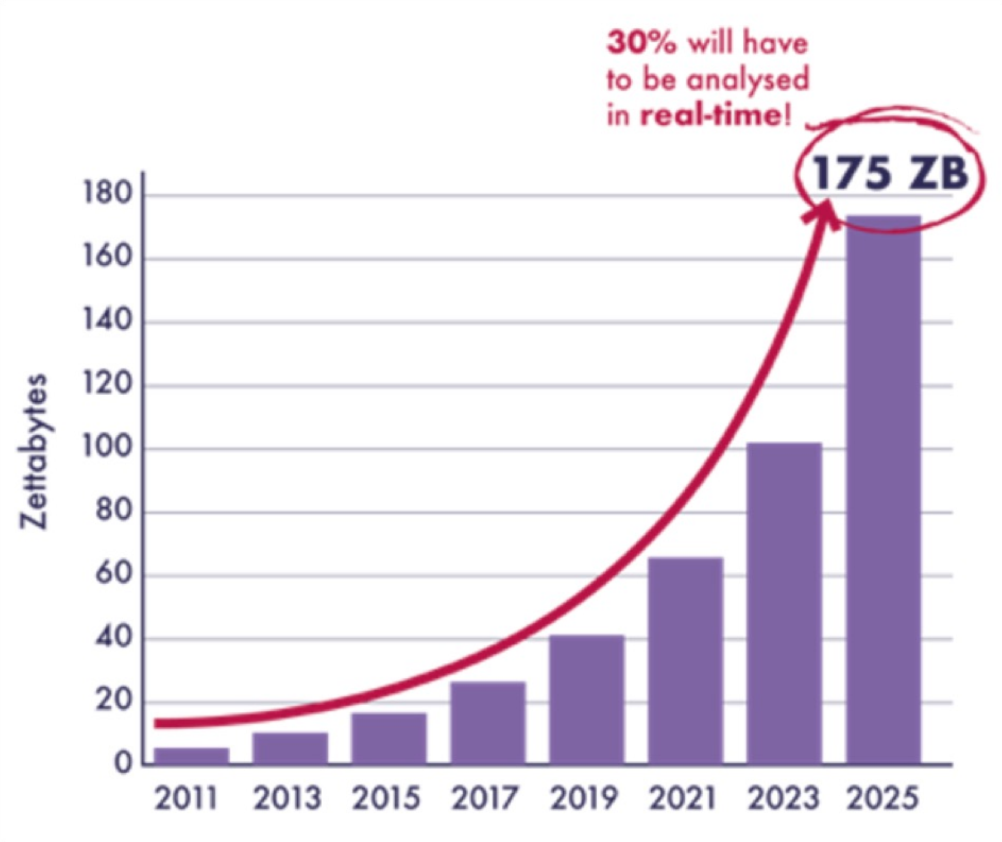
E. Della Valle, S. Ceri, F. van Harmelen, D. Fensel **It's a Streaming World! Reasoning upon Rapidly Changing Information**. IEEE Intelligent Systems 2009

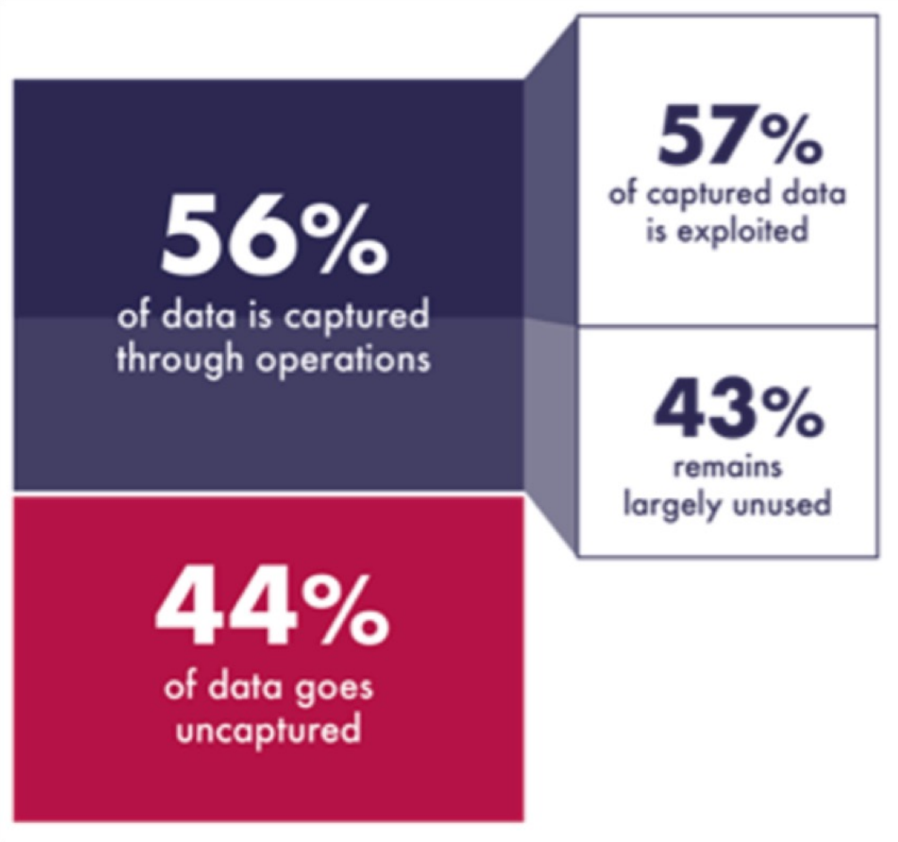
… but with conflicting requirements …

A system able to answer those queries must be able to

- handle volume
- handle velocity
- handle variety
- cope with incompleteness
- cope with noise
- provide reactive answers
- support fine-grained access
- integrate complex domain models
- offer high-level languages

… and 68% of data are not used



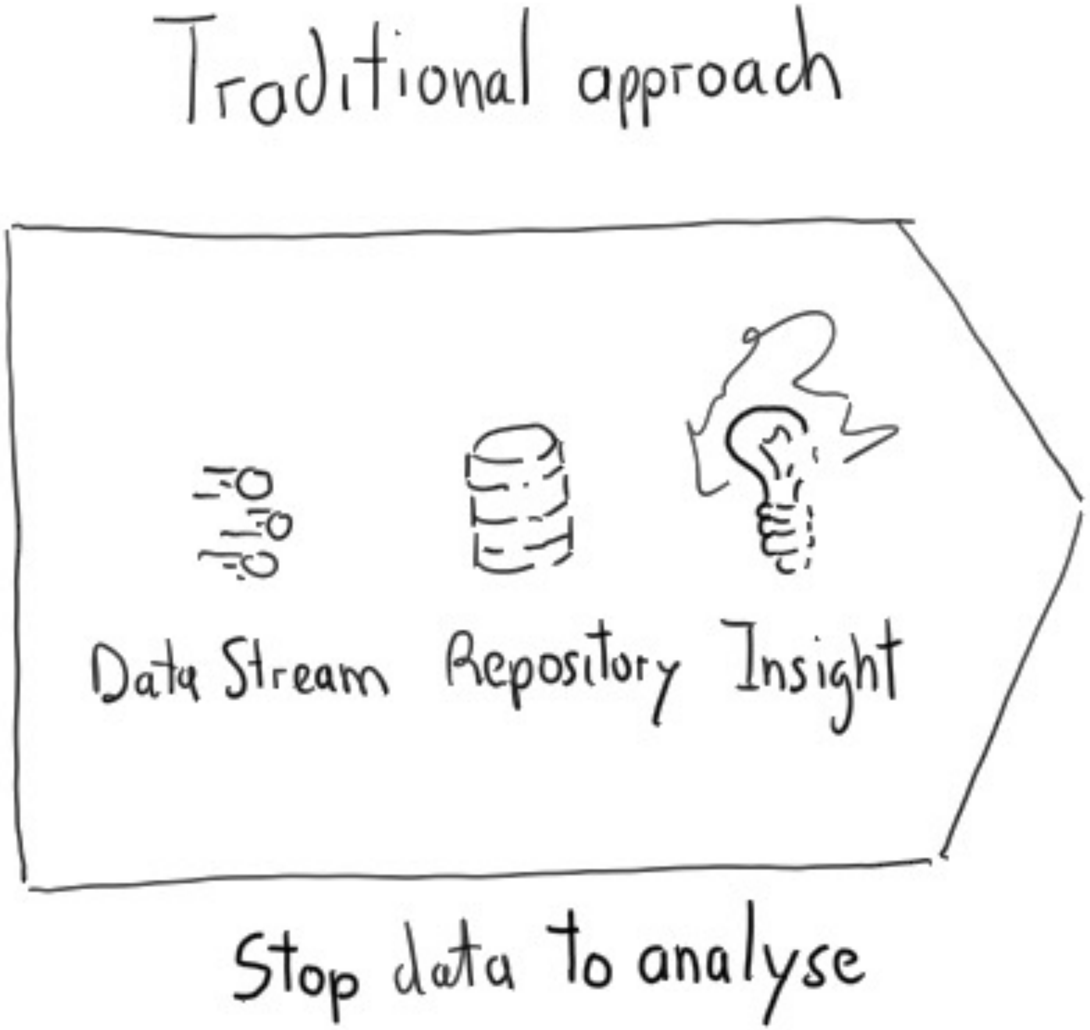


Src: Seagate

Internet minute in 2023

Src: eDiscovery Today & LTMG

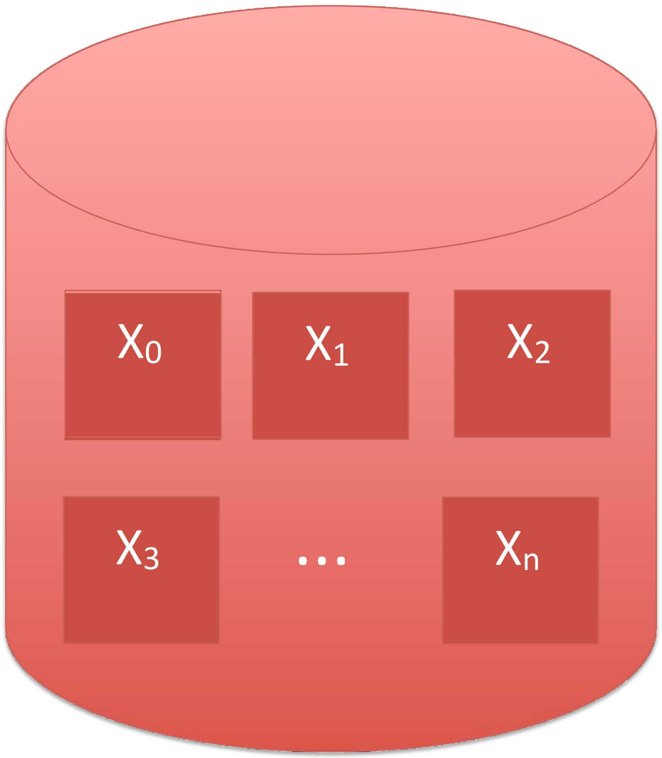
Traditional approach vs …



Batch of data

Random access
to data

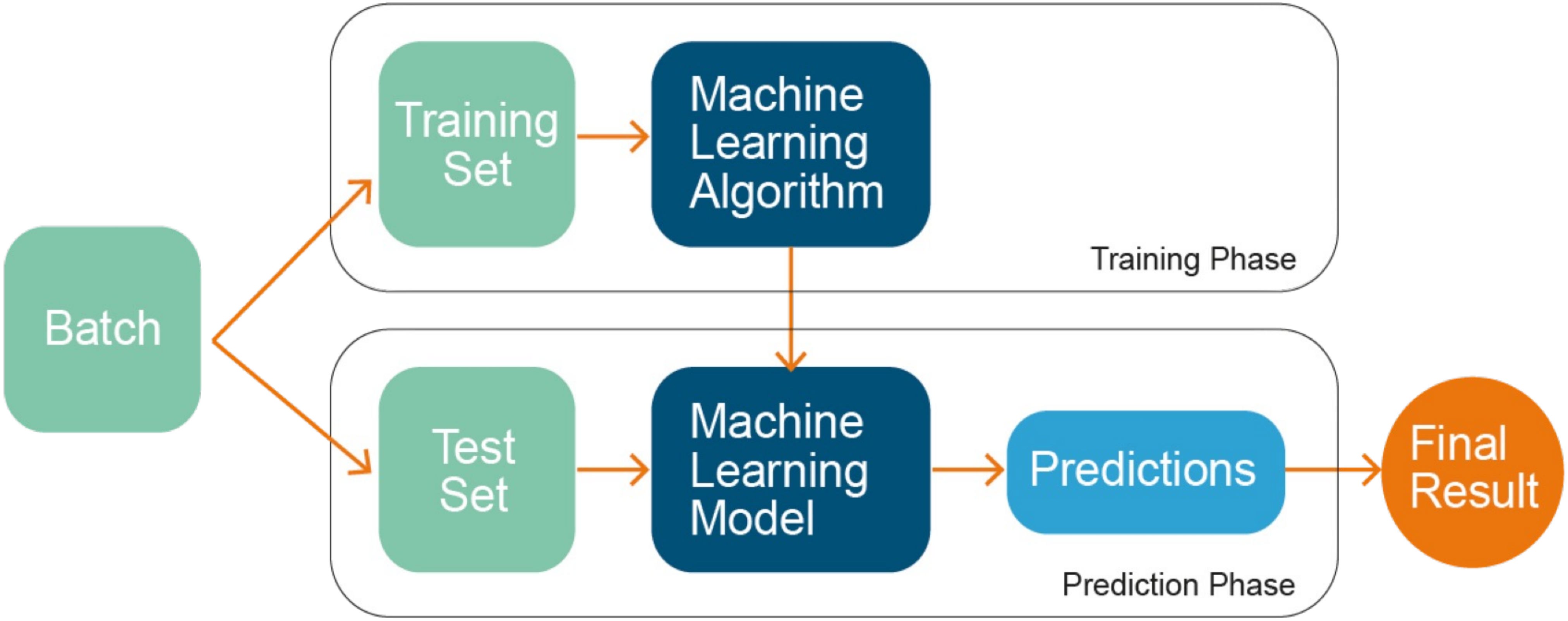
No restrictions on
memory/time
for training



Well defined
training phase

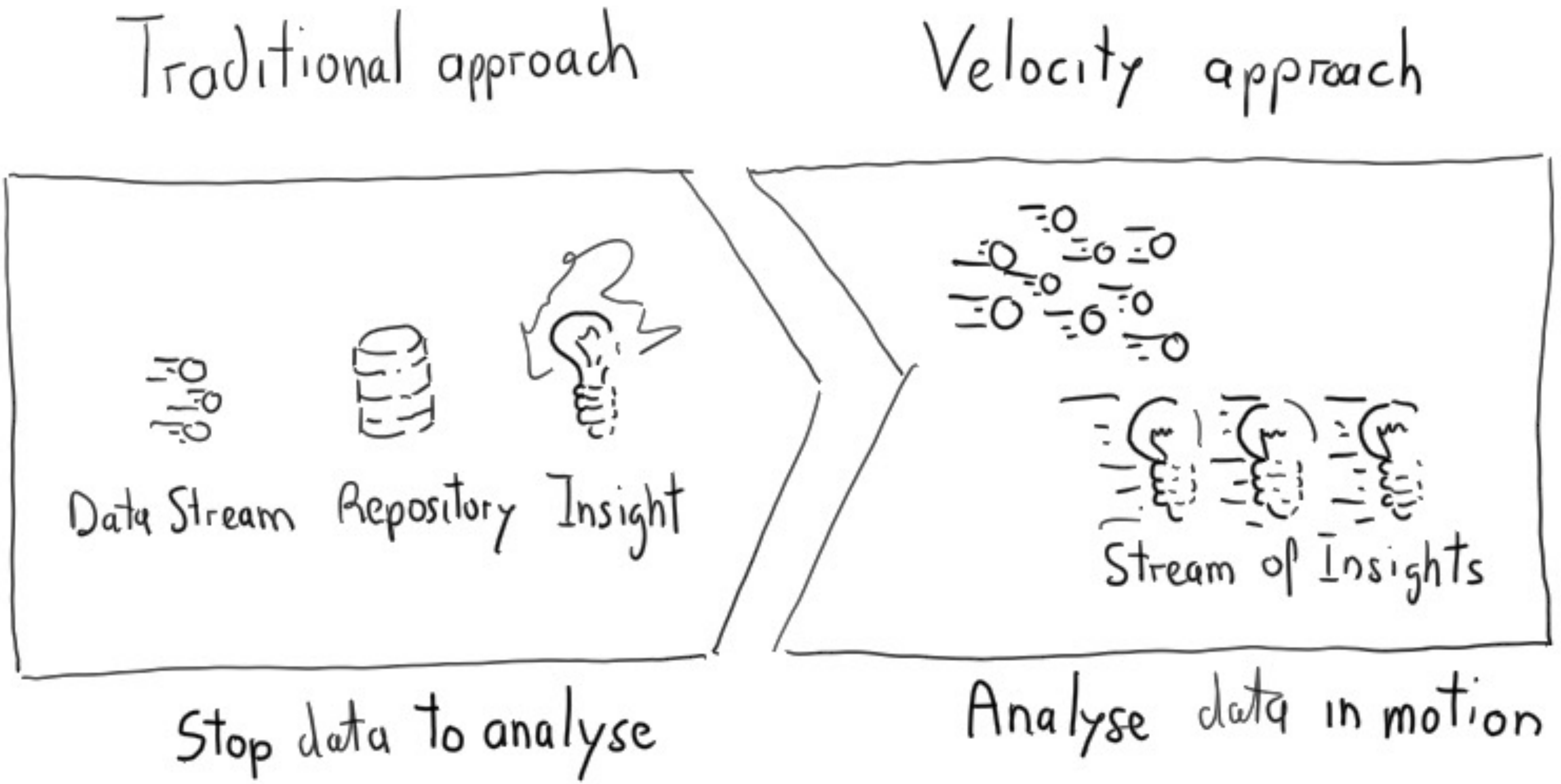
Access to all
labelled data
used for training

ML models



Batch
Training Set
Machine Learning Algorithm
Training Phase
Test Set
Machine Learning Model
Predictions
Final Result
Prediction Phase

Traditional vs. Velocity-oriented approach

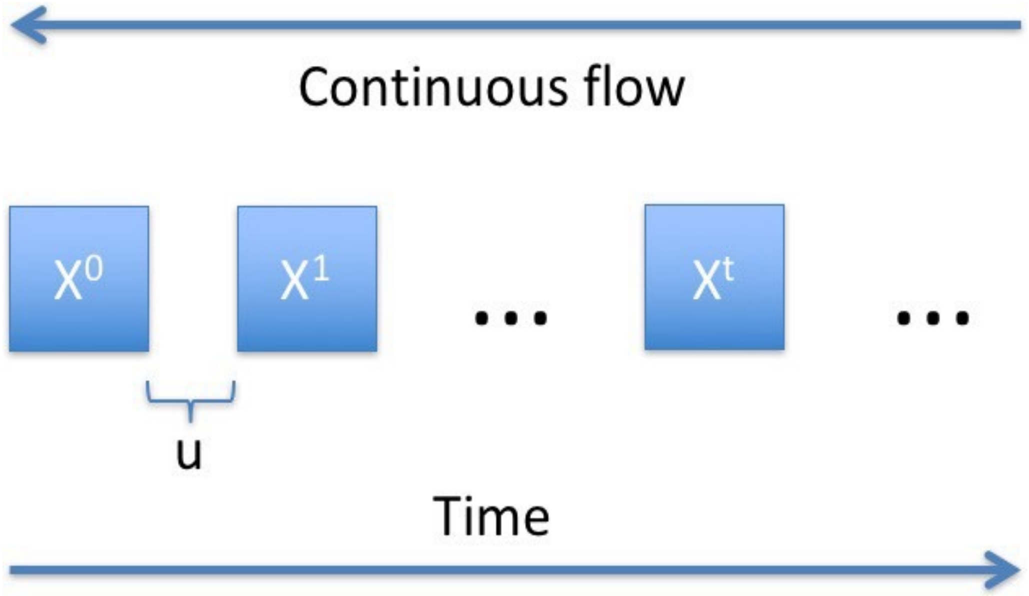


Data Stream

Continuous flow of data generated at **high-speed** in **dynamic**, **time-changing** environments

Sequential access only

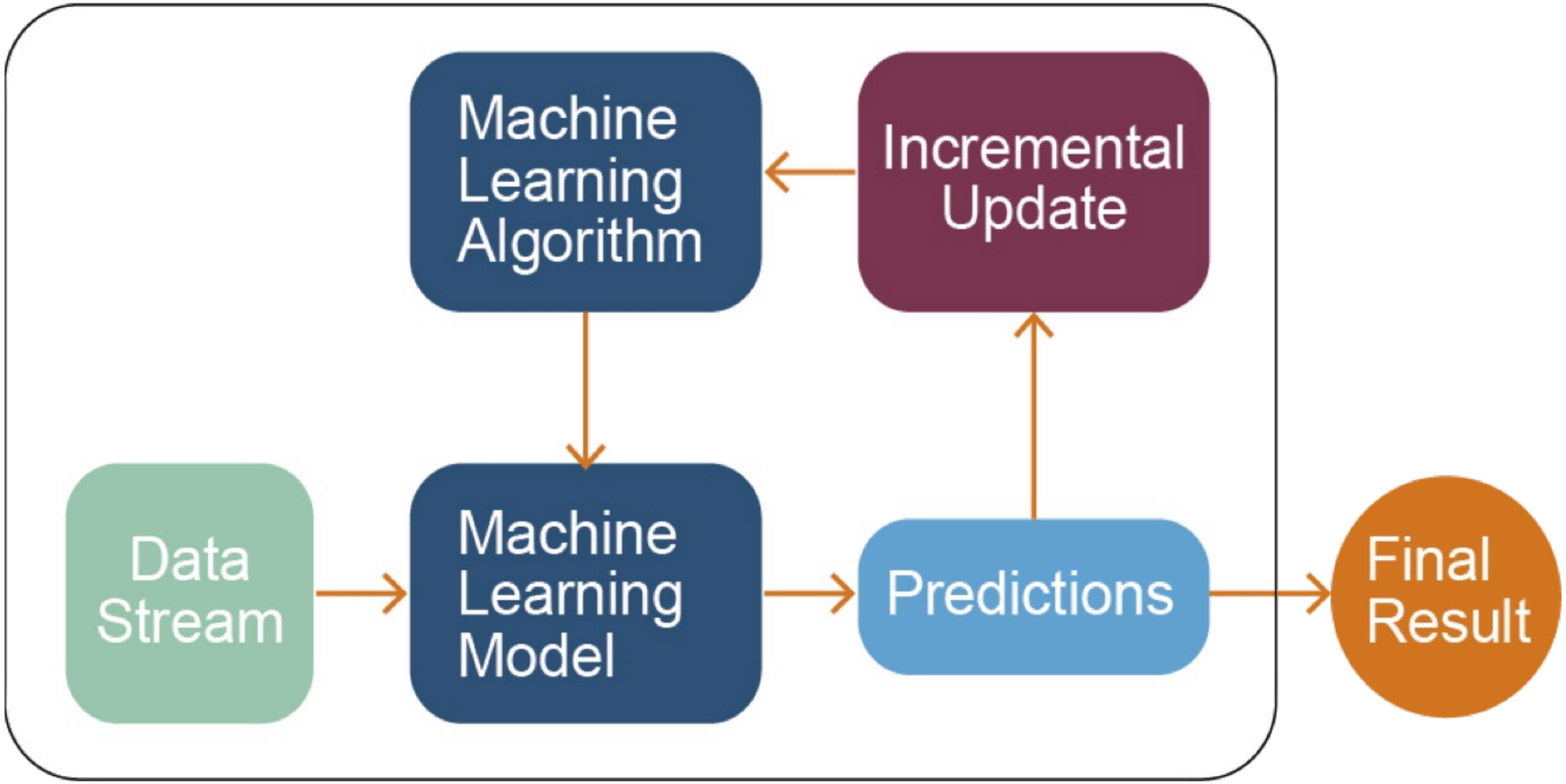
Evolving data



Strict time/memory requirements

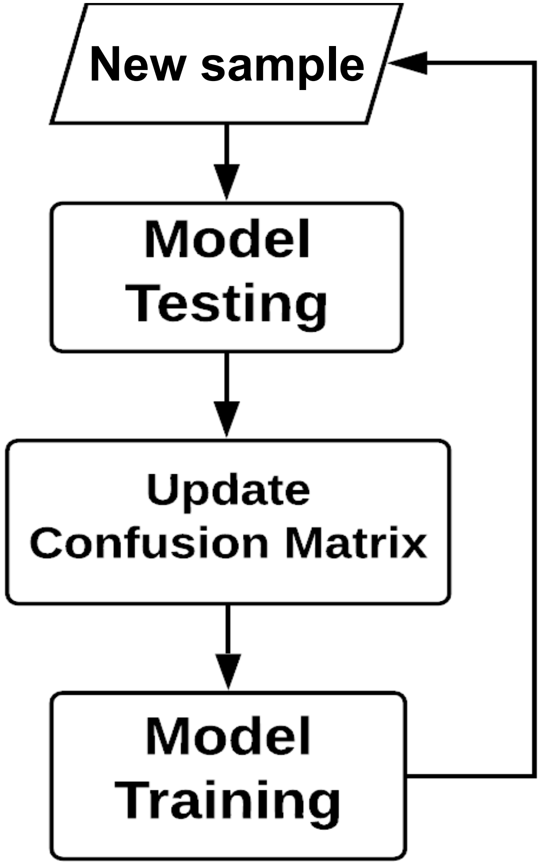
Characteristics of data seen so far

SML models



A. Bifet, G. de Francisci Morales, J. Read, G. Holmes, and B. Pfahringer **Efficient online evaluation of big data stream classifiers**. ACM SIGKDD 2015

Prequential evaluation



Estimate prequential error (PE):

- **Sliding window of size *w***

e_1	e_2	e_3	e_4	e_5		
	e_2	e_3	e_4	e_5	e_6	
		e_3	e_4	e_5	e_6	e_7

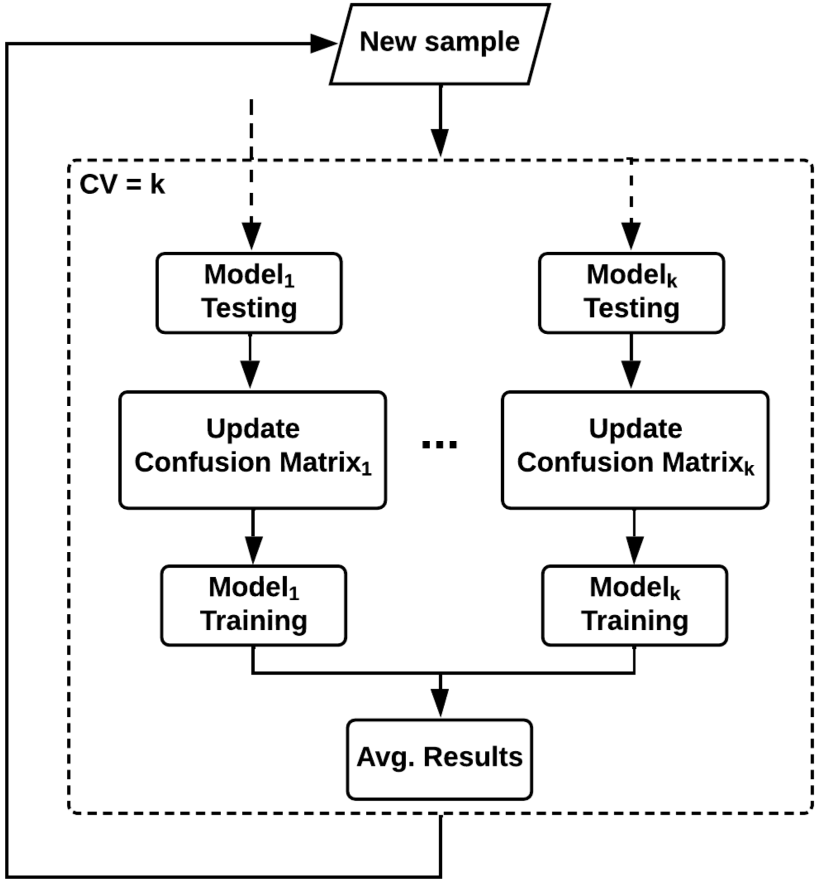
$$PE_i = \frac{1}{w} \sum_{k=i-w+1}^{w} e_k$$

- **Fading factor**

$$PE_i = \frac{\sum_{k=1}^{i} a^{i-k} * e_k}{\sum_{k=1}^{i} a^{i-k}} \qquad \text{with } 0 \ll \alpha \ \leq 1$$

Gama, J., Sebastião, R. and Rodrigues, P.P.: **Issues in evaluation of stream learning algorithms.** In ACM KDD, 2009.

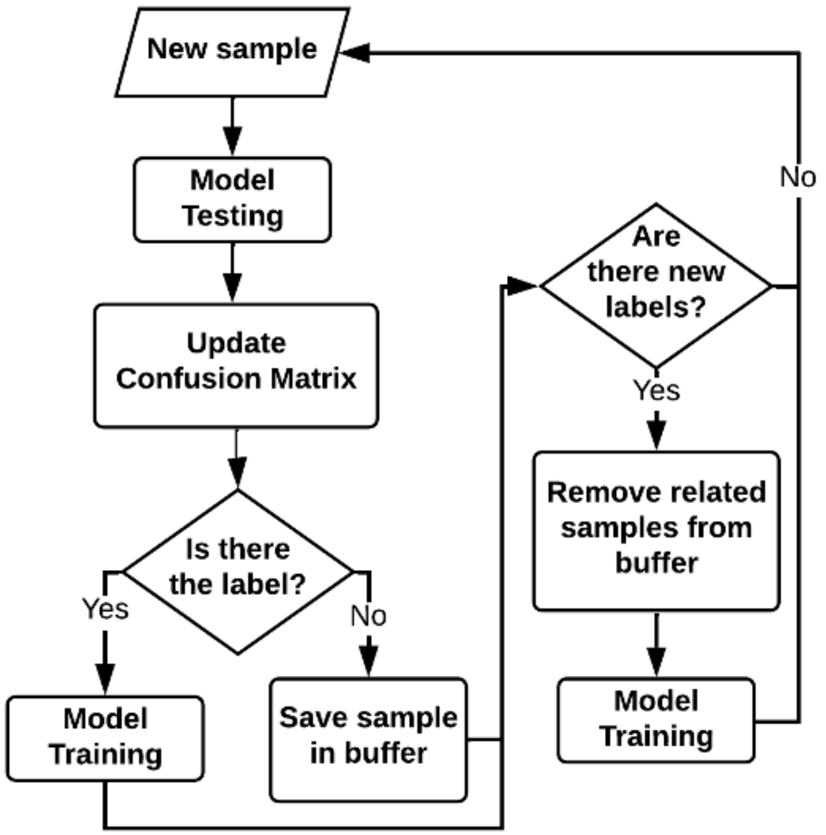
Prequential evaluation - Cross Validation



- **K-fold distributed cross-validation:**
  each sample is used for testing in one classifier selected randomly, and used for training and testing all the others
- **K-fold distributed split-validation:**
  each sample is used for training in one classifier selected randomly, and for testing in all the classifiers
- **K-fold distributed bootstrap-validation:**
  each sample is used for training in approximately 2⁄3 of the classifiers, with a separate weight in each classifier, and for testing in all the classifiers

Bifet, A., et al: **Efficient Online Evaluation of Big Data Stream Classifiers**. In ACM SIGKDD, 2015.

Prequential evaluation - Delayed



- In real environments, can happen that the label arrives **delayed** w.r.t. the features
- Test the model with the features and wait for the label to train it

Gomes, HM., et al: **Adaptive random forests for evolving data stream classification**. In Machine Learning, 2017.

Evaluation metric - Kappa statistic

$$k = \frac{p - p_{rand}}{1 - p_{rand}}$$

where p is the accuracy of the classifier under consideration and p_{rand} is the accuracy of the Random classifier.

- If the classifier is perfectly correct, then $k = 1$.
- If the classifier achieves the same accuracy as the Random classifier, then $k = 0$.

I. Žliobaitė et al. **Evaluation methods and decision theory for classification of streaming data with temporal dependence.** In Machine Learning, 2015.

Evaluation metric - Kappa-Temporal statistic

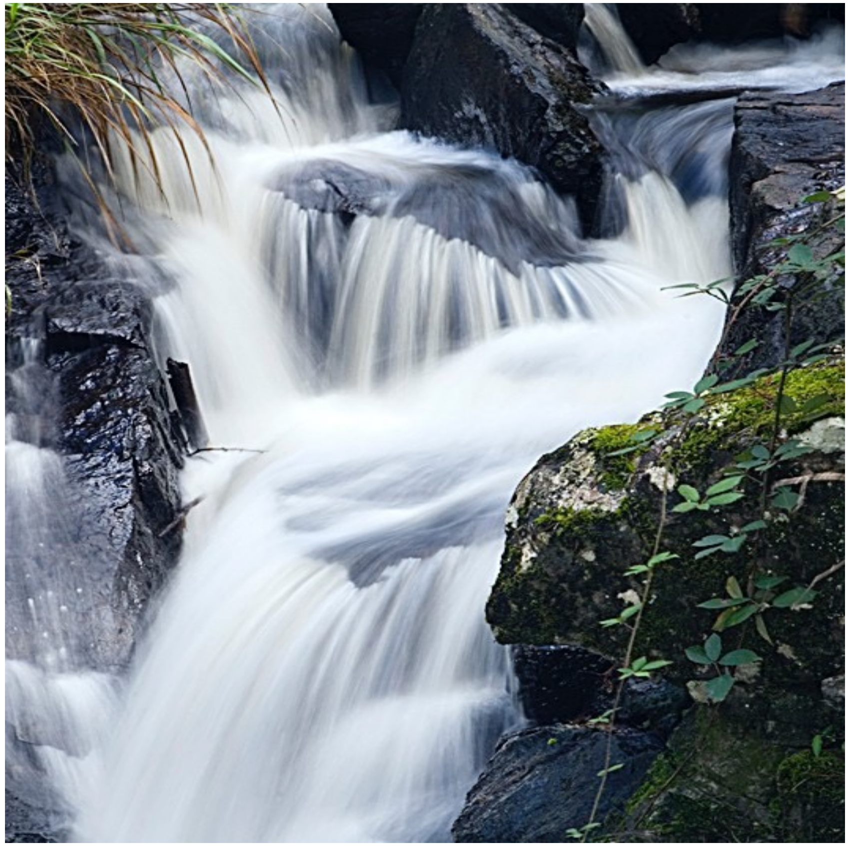
$$k = \frac{p - p_{per}}{1 - p_{per}}$$

where p is the accuracy of the classifier under consideration and p_{per} is the accuracy of the Persistent classifier.

- If the classifier is perfectly correct, then $k = 1$.
- If the classifier achieves the same accuracy as the Persistent classifier, then $k = 0$.
- If the classifier performs worse than the Persistent classifier, then $k < 0$.

I. Žliobaitè et al. **Evaluation methods and decision theory for classification of streaming data with temporal dependence.** In Machine Learning, 2015.

SML models



- Incrementally incorporate data on the fly
- Unbounded real-time data
- Resource efficient
- Dynamic models

SML models

Benefits

- One sample at a time
- Incremental models
- Time & Memory management

Challenges

- Data stationarity (concept drift)
- Class imbalance
- Hyper-parameter tuning

POLITECNICO
MILANO 1863

Exercise 1: From batch to stream learning

Credits

- Albert Bifet DATA STREAM MINING 2020-2021 course at Telecom Paris
- Alessio Bernardo & Emanuele Della Valle

Streaming Machine Learning Introduction

Alessio Bernardo

Post-doc @ Politecnico di Milano

CTO & Co-founder @ Motus ml

POLITECNICO
MILANO 1863