

Streaming Machine Learning (SML)

Alessio Bernardo & Emanuele Della Valle
05-06-2021

About me



Alessio Bernardo

Ph.D. Student in Data Science:

- Politecnico di Milano
- Research on Streaming Machine Learning

M.Sc. & B.Sc. Computer Engineering:

- Politecnico di Milano

Part I

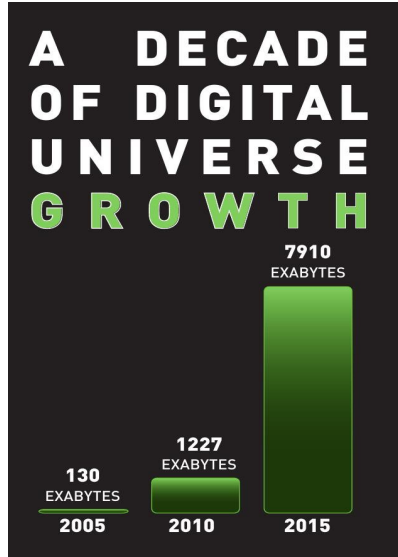
Introduction

Credits

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

Big Data Trend

Big Data Trend



The **amount of information managed** by enterprise datacenters will grow by **50 times.**

Meanwhile, the **number of IT professionals** in the world will grow by **less than 1.5 times.**

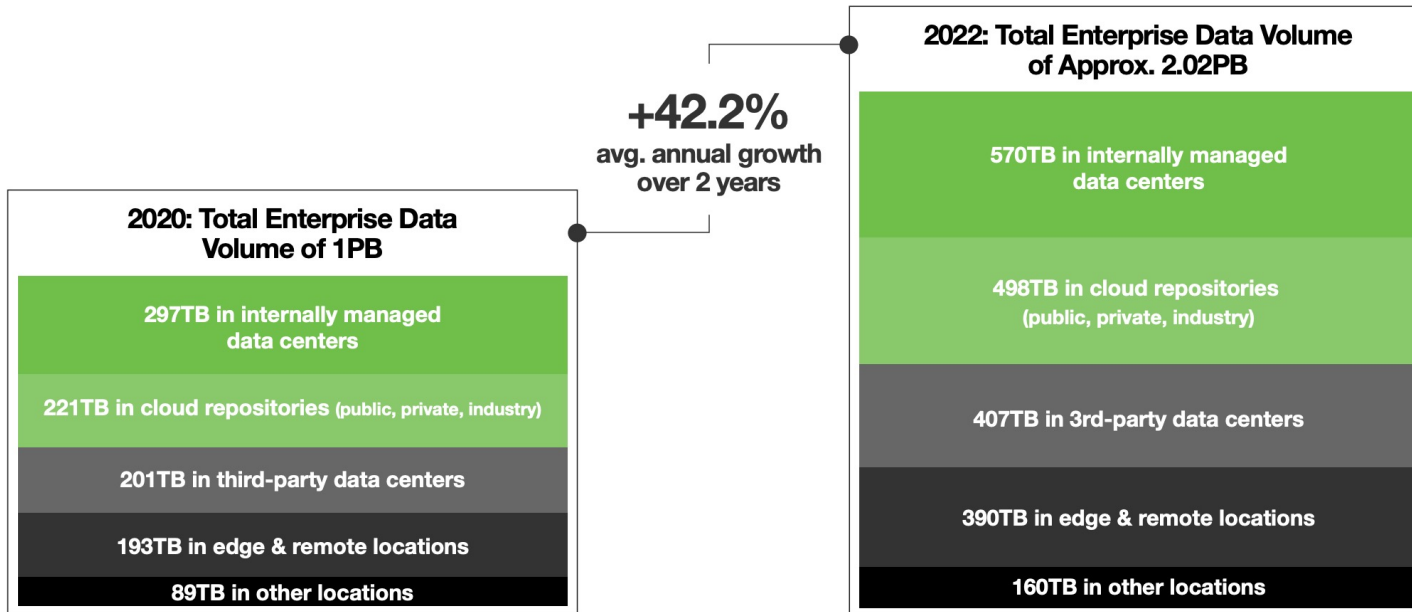
Source: IDC's Digital Universe Study (EMC), June 2011

Big Data Trend



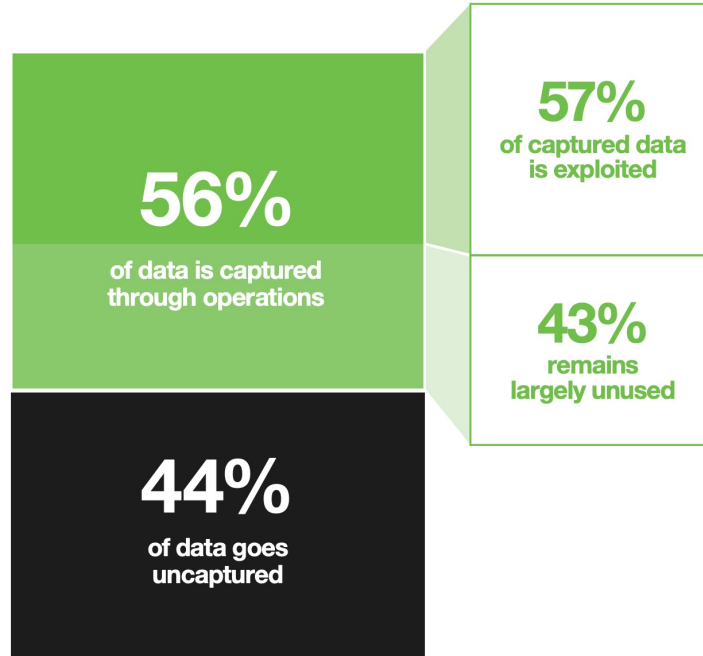
Source: @LoriLewis and @OfficiallyChadd

Big Data Trend



Source: SEAGATE TECHNOLOGY, June 2020

Big Data Trend



Source: SEAGATE TECHNOLOGY, June 2020

Big Data Trend

Data is growing, and the rate of growth is accelerating. The sum of data generated by 2025 is set to accelerate exponentially to 175 zettabytes, an order of magnitude bigger than the storage production capability.

Dave Mosley,
CEO of SEAGATE TECHNOLOGY

Big Data Trend

*Innovation is **not** driven **by trends**, but **by the need** to create **more value** under constraints. This exponential inflation will thus require **analysing** almost **30%** of global data in **real-time**.*

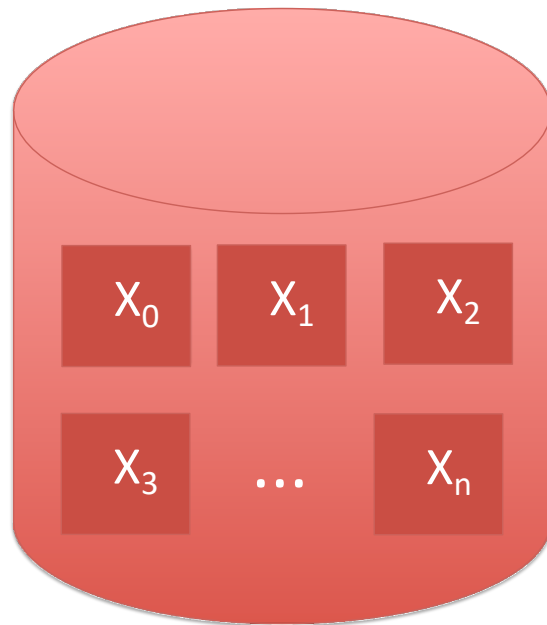
Dave Mosley,
CEO of SEAGATE TECHNOLOGY

Batch vs Data Stream

Batch

Random access
to data

No restrictions on
memory/time for
training

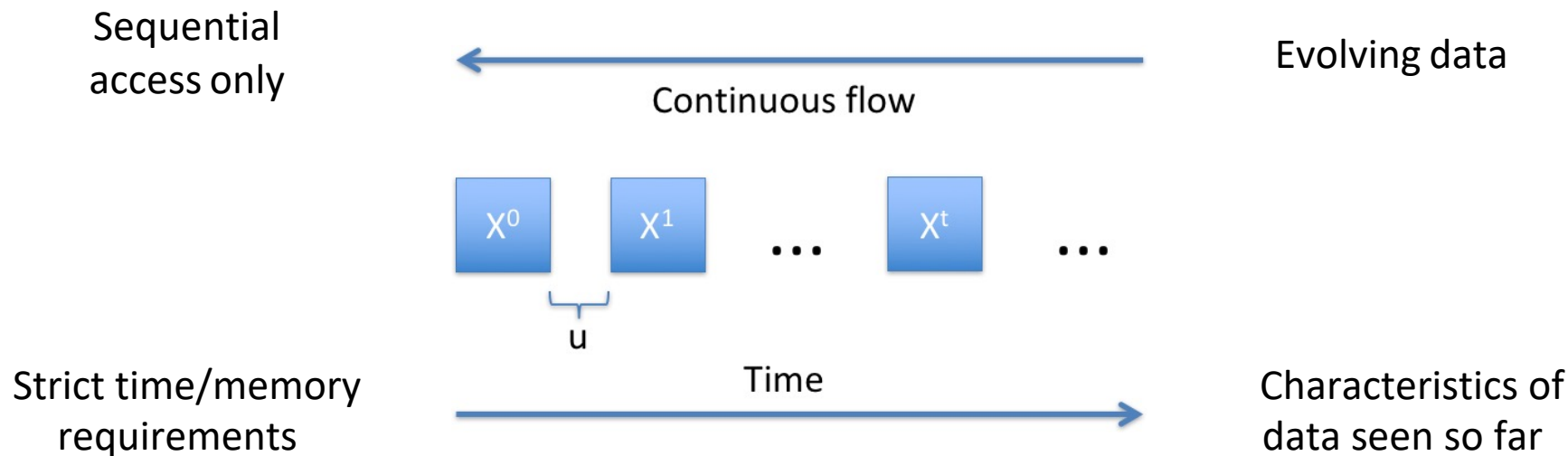


Well defined
training phase

Access to all labeled
data used for training

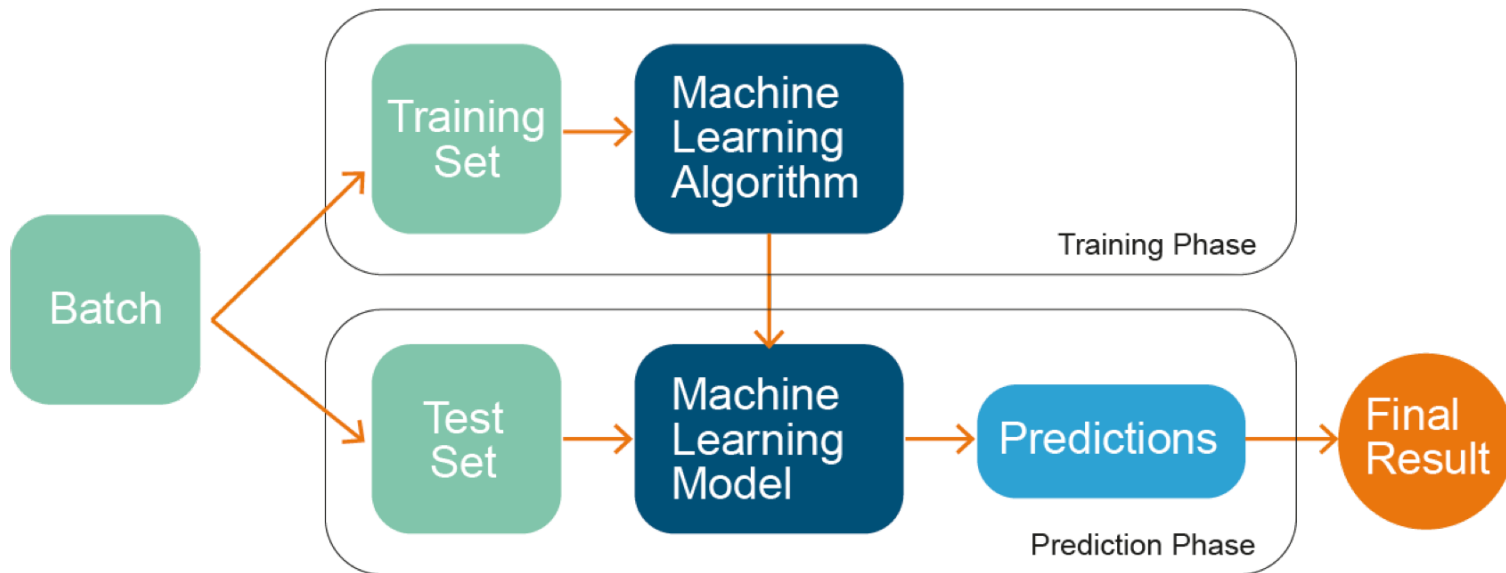
Data Stream

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

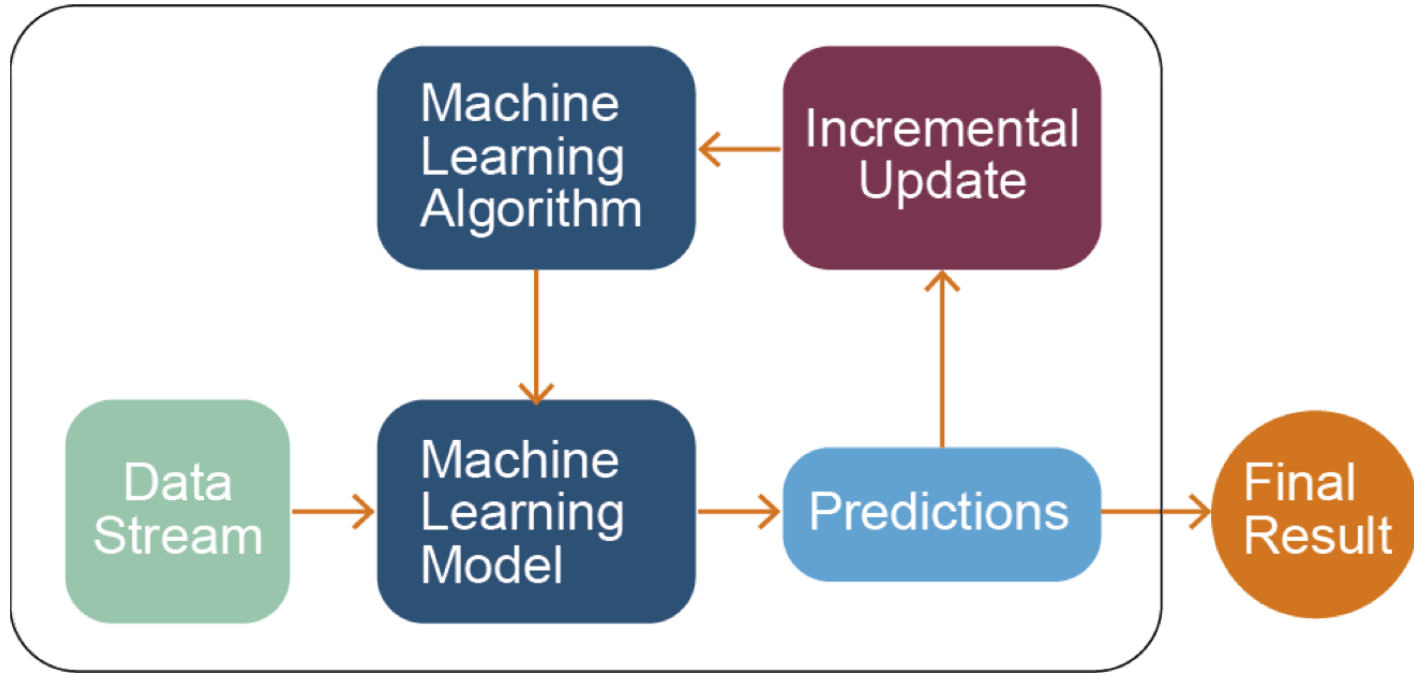


ML vs SML

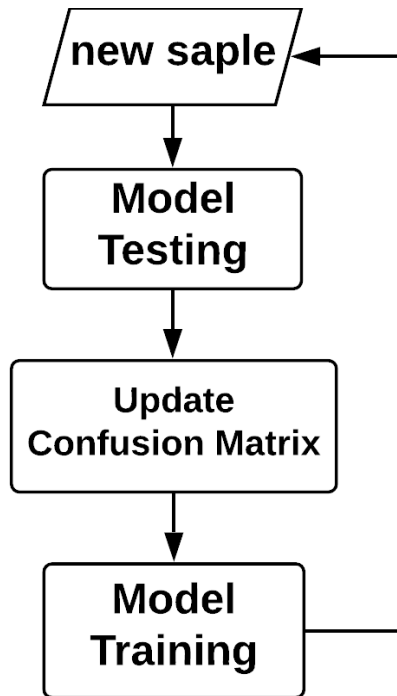
ML Models



SML Models

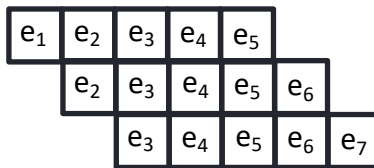


Prequential Evaluation



Estimate prequential error (PE):

- Sliding window of size w



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

- Fading factor

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

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

SML Models

- Incorporate data on the fly
- Unbounded training sets
- Resource efficient
- Dynamic models



Benefits

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

Challenges

- Non-stationarity (Concept drift)
- Class imbalance
- Hyper-parameter Tuning

QUIZ

1. What are the data streams characteristics?
 - a. All data are available, non-stationary, bounded
 - b. One sample available at a time, non-stationary, unbounded
 - c. One sample available at a time, unbounded, access to old data
2. How do the SML models address the time and memory problem?
 - a. Updating the model with the new sample and then discarding it
 - b. Updating the model with the new sample and then saving it
 - c. Saving every time the new sample and retraining anew the model

QUIZ

1. What are the data streams characteristics?
 - a. All data are available, non-stationary, bounded
 - b. **One sample available at a time, non-stationary, unbounded**
 - c. One sample available at a time, unbounded, access to old data
2. How do the SML models address the time and memory problem?
 - a. Updating the model with the new sample and then discarding it
 - b. Updating the model with the new sample and then saving it
 - c. Saving every time the new sample and retraining anew the model

QUIZ

1. What are the data streams characteristics?
 - a. All data are available, non-stationary, bounded
 - b. **One sample available at a time, non-stationary, unbounded**
 - c. One sample available at a time, unbounded, access to old data
2. How do the SML models address the time and memory problem?
 - a. **Updating the model with the new sample and then discarding it**
 - b. Updating the model with the new sample and then saving it
 - c. Saving every time the new sample and retraining anew the model

EXERCISE 1: From batch to stream learning

LAB 1: Prequential error
