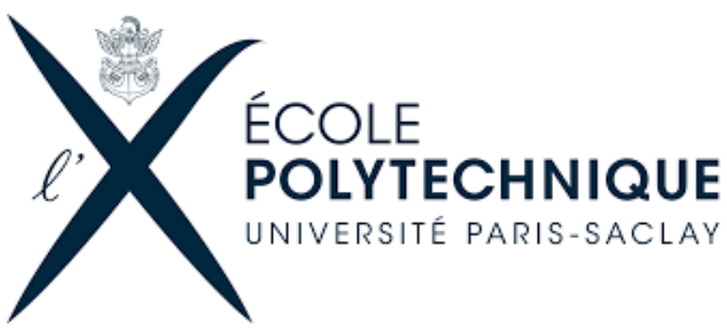
Machine Learning for Data Streams

Albert Bifet (@abifet)

DATAIA-JST International
Symposium on Data Science and AI
11 July 2018

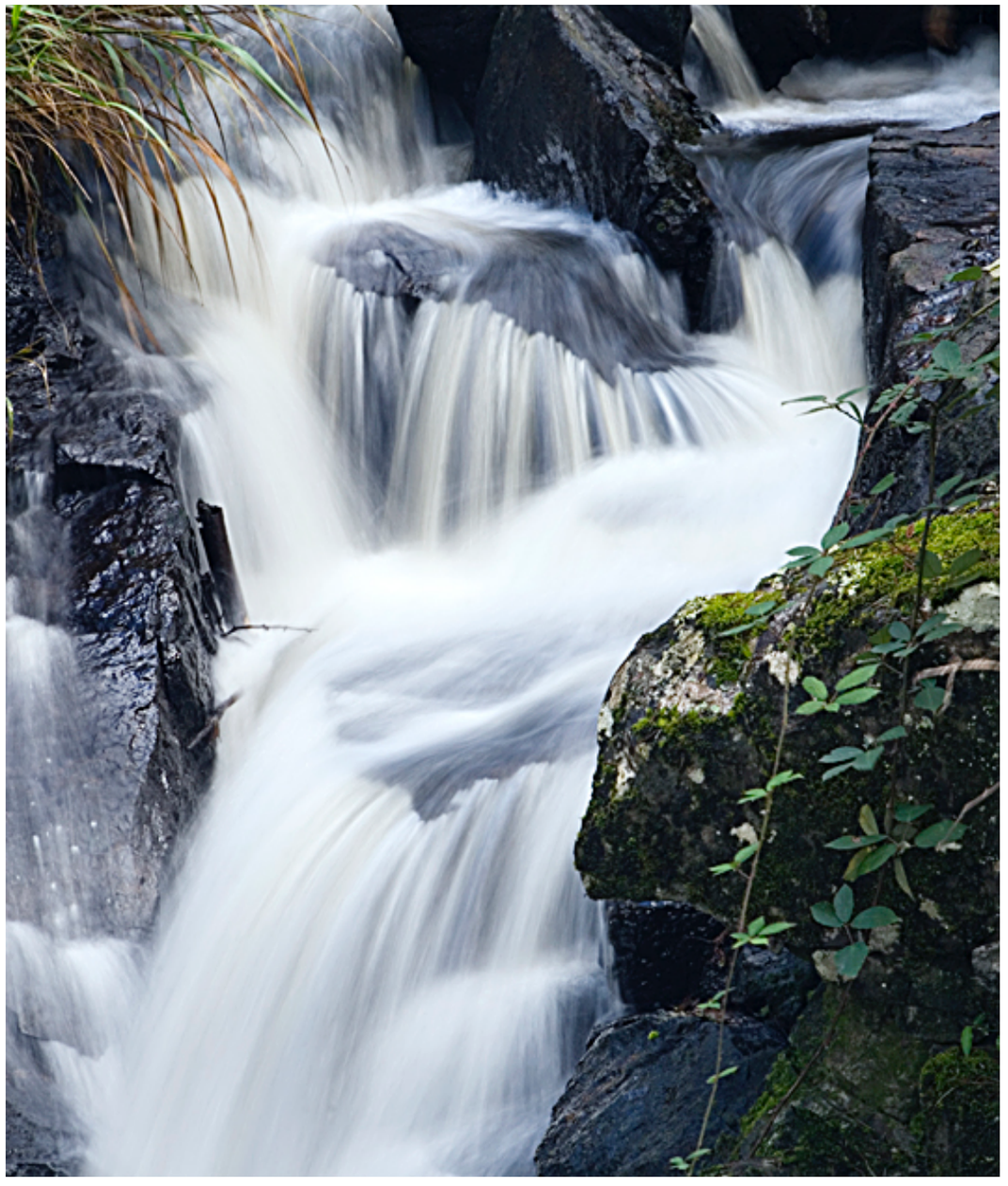


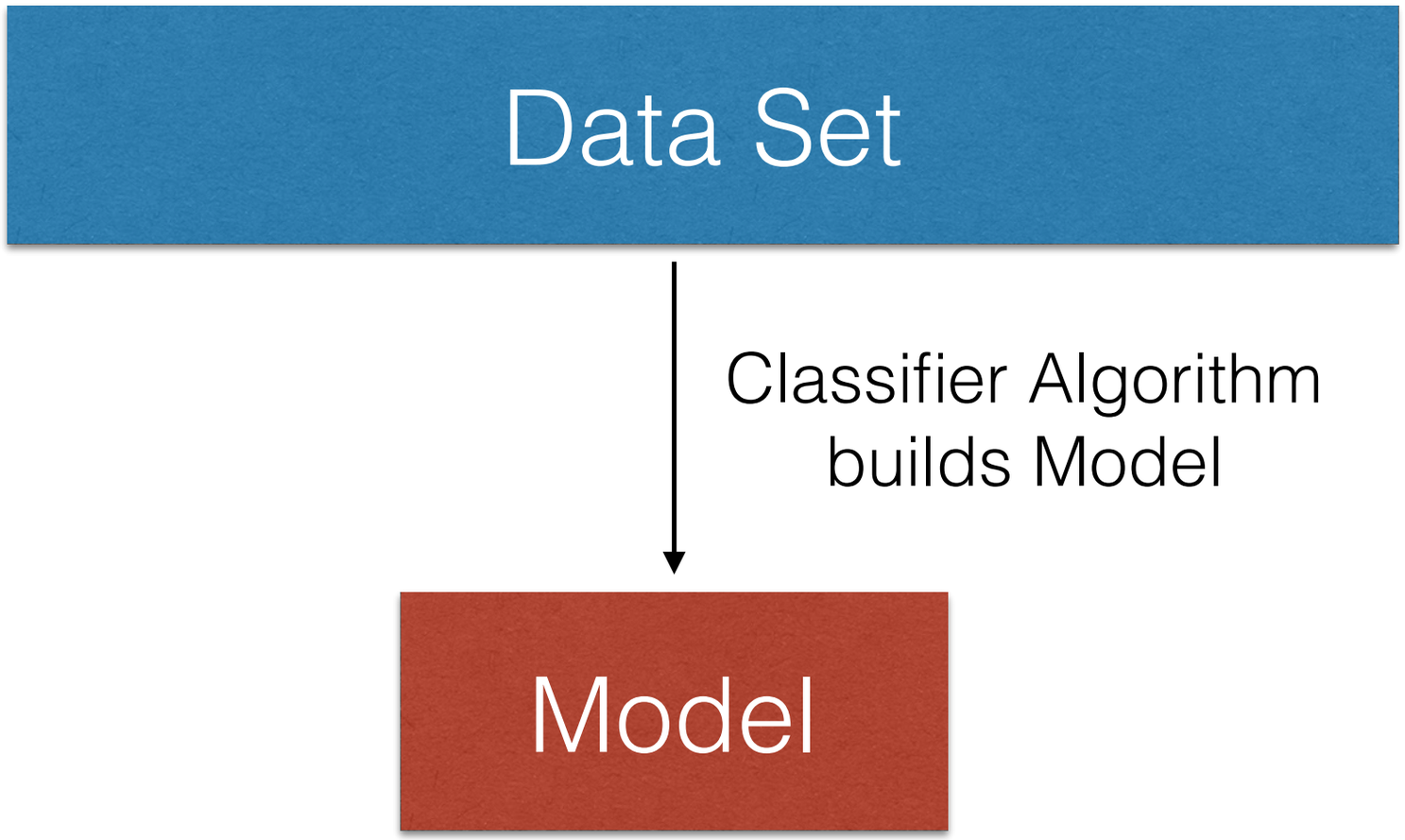
AI Systems

- According to **Nikola Kasabov,** AI systems should exhibit the following characteristics:
  - Accommodate new problem solving rules **incrementally**
  - **Adapt online and in real time**
  - Are able to **analyze itself** in terms of behavior, error and success.
  - Learn and improve through interaction with the environment (embodiment)
  - Learn quickly from large amounts of data (**Big Data**)
  - Have memory-based exemplar storage and retrieval capacities
  - Have parameters to represent short and long term memory, age, forgetting, etc.

Data Streams

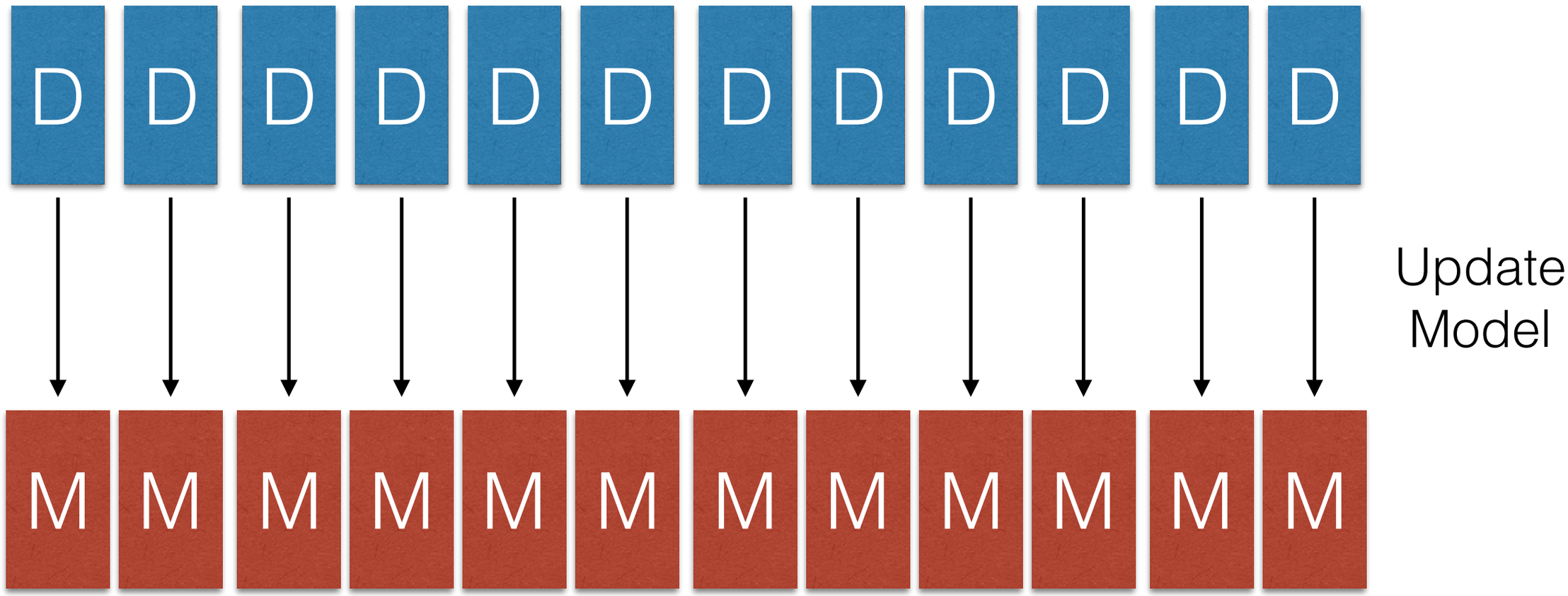
- Maintain models online
  - Incorporate data on the fly
  - Unbounded training sets
  - Resource efficient
  - Detect changes and adapts
  - Dynamic models





Analytic Standard Approach

Finite training sets
Static models

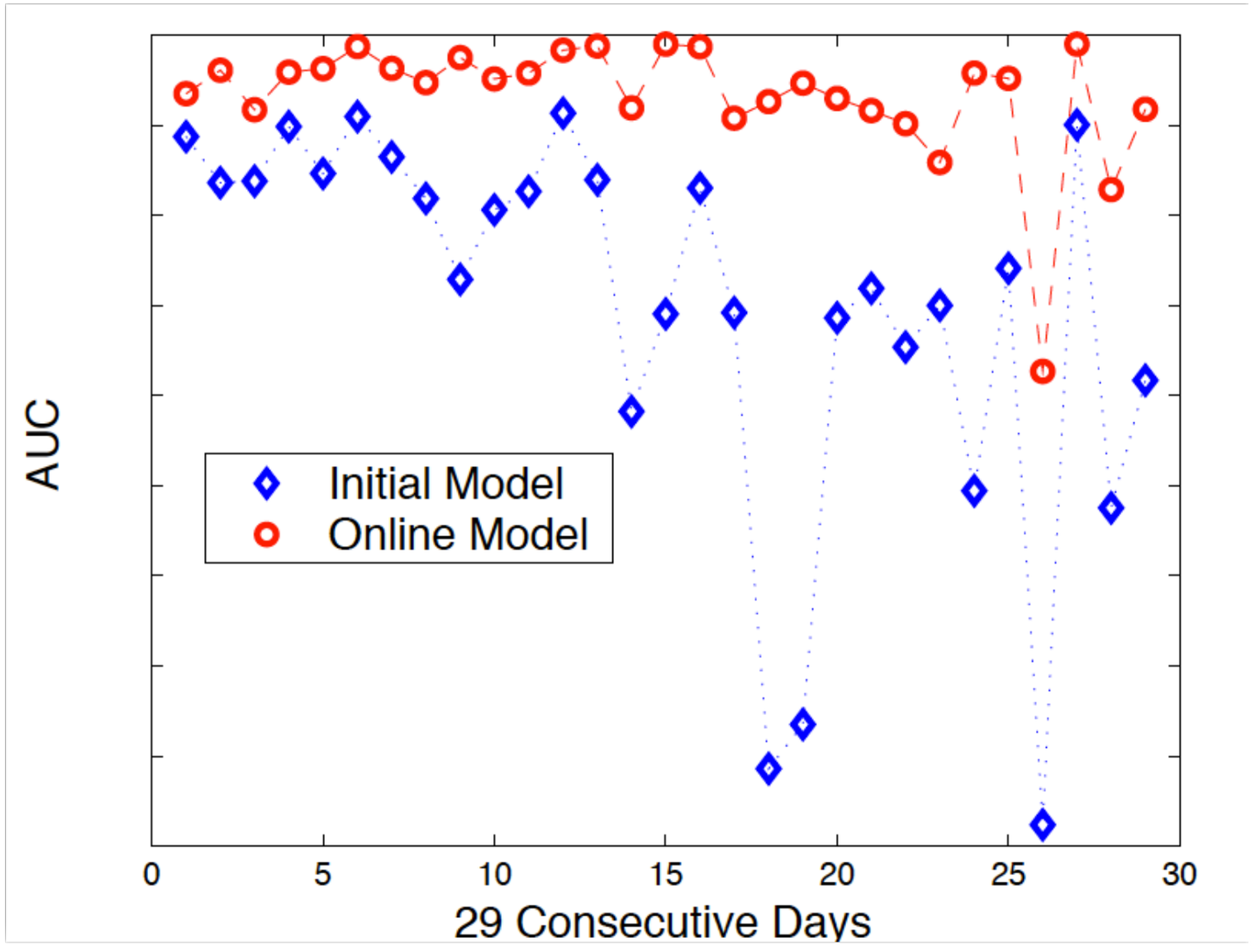


Data Stream Approach

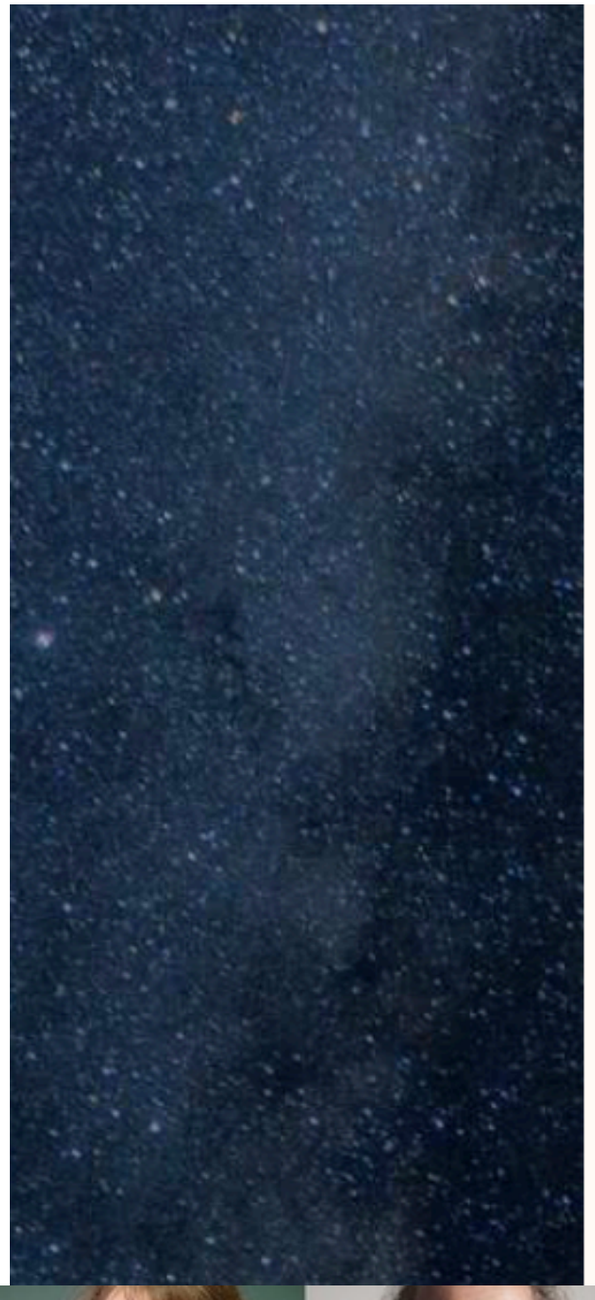
Infinite training sets
Dynamic models

Adversarial Learning

- Need to **retrain!**
  - Things change over time
  - How often?
- Data unused until next update!
  - Value of data wasted



AI Challenges



Cédric Villani and Marc Shoenauer

CÉDRIC VILLANI

Mathematician and
Member of the French Parliament

FOR A MEANINGFUL ARTIFICIAL INTELLIGENCE

TOWARDS A FRENCH
AND EUROPEAN STRATEGY

1. Green AI

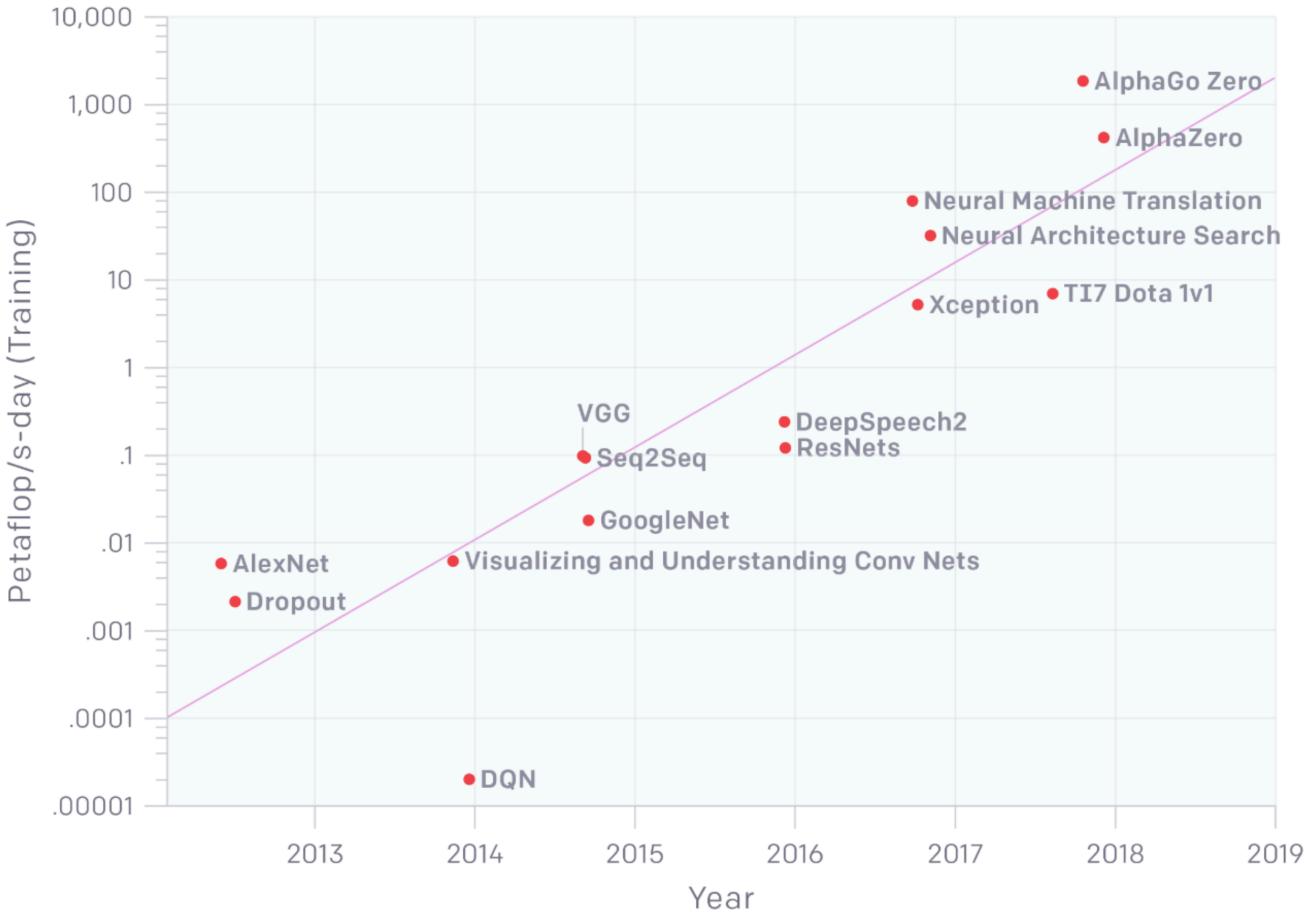
Part 4 — Using Artificial Intelligence to Help Create a More Ecological Economy

More than ever before, the revolution triggered by the development of digital technologies and their widespread adoption tends to obscure its impact on the environment[1]. Nevertheless, there is an urgent need to take this on board. Two years ago, the American Association of Semi-Conductor Manufacturers predicted that by 2040, the global demand for data storage capacity, which grows at the pace of the progress of AI, will exceed the available world production of silicon[2].

By 2040 the energy required for computation will equally have exceeded world energy production

Furthermore, by 2040 the energy required for computation will equally have exceeded world energy production; the progress of the blockchain may also cause our energy requirements to rocket. It is vital to educate as many people as possible about these issues and to act promptly to avoid shortages. At a time when global warming is a scientific certainty, it is no longer possible to pursue technological and societal developments if those are completely detached from the need to preserve our environment.

AlexNet to AlphaGo Zero: A 300,000x Increase in Compute



Green AI

- One pass over the data
- Approximation algorithms: small error ε with high probability 1-δ
  - True hypothesis H, and learned hypothesis $\hat{H}$
  - $Pr[\ |H - \hat{H}| < \varepsilon |H|\ ] > 1-\delta$

2. Explainable AI

relations and reinforce solidarity. Diversity should also figure within these priorities. In this respect, the situation in the digital sector is alarming, with women very poorly represented. Their under-representation may lead to the spread of nurture gender-biased algorithms.

Our digital society cannot be governed by black box algorithms

Finally, our digital society could not be governed by black box algorithms: artificial intelligence is going to play a decisive role in critical domains for human flourishing (health, banking, housing, etc) and there is currently a high risk of embedding existing discrimination into AI algorithms or creating new areas where it might occur. Further, we also run the risk that normalization may spread attitudes that could lead to the general development of algorithms within artificial intelligence. It should be possible to open these black boxes, but equally to think ahead about the ethical issues that may be raised by algorithms within artificial intelligence.

A meaningful AI finally implies that AI should be explainable: explaining this technology to the public so as to demystify it—and the role of the media is vital from this point of view—but also explaining artificial intelligence by extending research into explicability itself. AI specialists themselves frequently maintain that significant advances could be made on this subject.

Pedro Domingos
@pmddomingos

Starting May 25, the European Union will require algorithms to explain their output, making deep learning illegal.

12:59 AM - Jan 29, 2018

343 248 people are talking about this

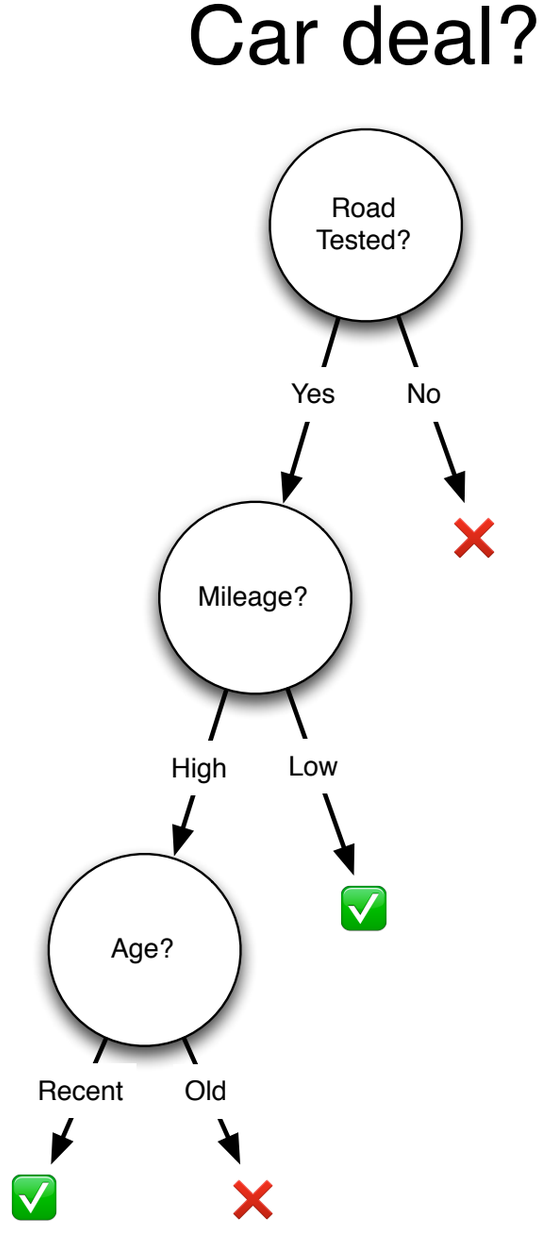
Art. 22 GDPR

Automated individual decision-making, including profiling

(1) The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her.

Decision Tree

- Each node tests a features
- Each branch represents a value
- Each leaf assigns a class
- Greedy recursive induction
  - Sort all examples through tree
  - x_i = most discriminative attribute
  - New node for x_i, new branch for each value, leaf assigns majority class
  - Stop if no error | limit on #instances



P. Domingos and G. Hulten, "Mining High-Speed Data Streams," KDD '00

HOEFFDING TREE

- Sample of stream enough for near optimal decision
- Estimate merit of alternatives from prefix of stream
- Choose sample size based on statistical principles
- When to expand a leaf?
  - Let x_1 be the most informative attribute, x_2 the second most informative one
  - Hoeffding bound: split if $G(x_1) - G(x_2) > \boldsymbol{\varepsilon} = \sqrt{\frac{R^2 \ln(1/\delta)}{2n}}$

TensorForest: Scalable Random Forests on TensorFlow

Thomas Colthurst, Gilbert Hendry, Zachary Nado, D. Sculley
Google Inc.
{thomaswc, gilberth, znado, dsculley}@google.com

Abstract

We present TensorForest, a highly scalable open-sourced system built on top of TensorFlow for the training and evaluation of random forests. TensorForest achieves scalability by combining a variant of the online Hoeffding Tree algorithm with the extremely randomized approach, and by using TensorFlow's native support for distributed computation. This paper describes TensorForest's architecture, analyzes several alternatives to the Hoeffding bound for per-node split determination, reports performance on a selection of large and small public datasets, and demonstrates the benefit of tight integration with the larger TensorFlow platform.

Adaptive Random Forest

- Why Random Forests?
  - Off-the-shelf learner
  - Good learning performance

Adaptive random forests for evolving data stream classification.

Gomes, H M; Bifet, A; Read, J; Barddal, J P; Enembreck, F; Pfharinger, B; Holmes, G; Abdessalem, T.

Machine Learning, Springer, 2017.

- Based on the original Random Forest by Breiman

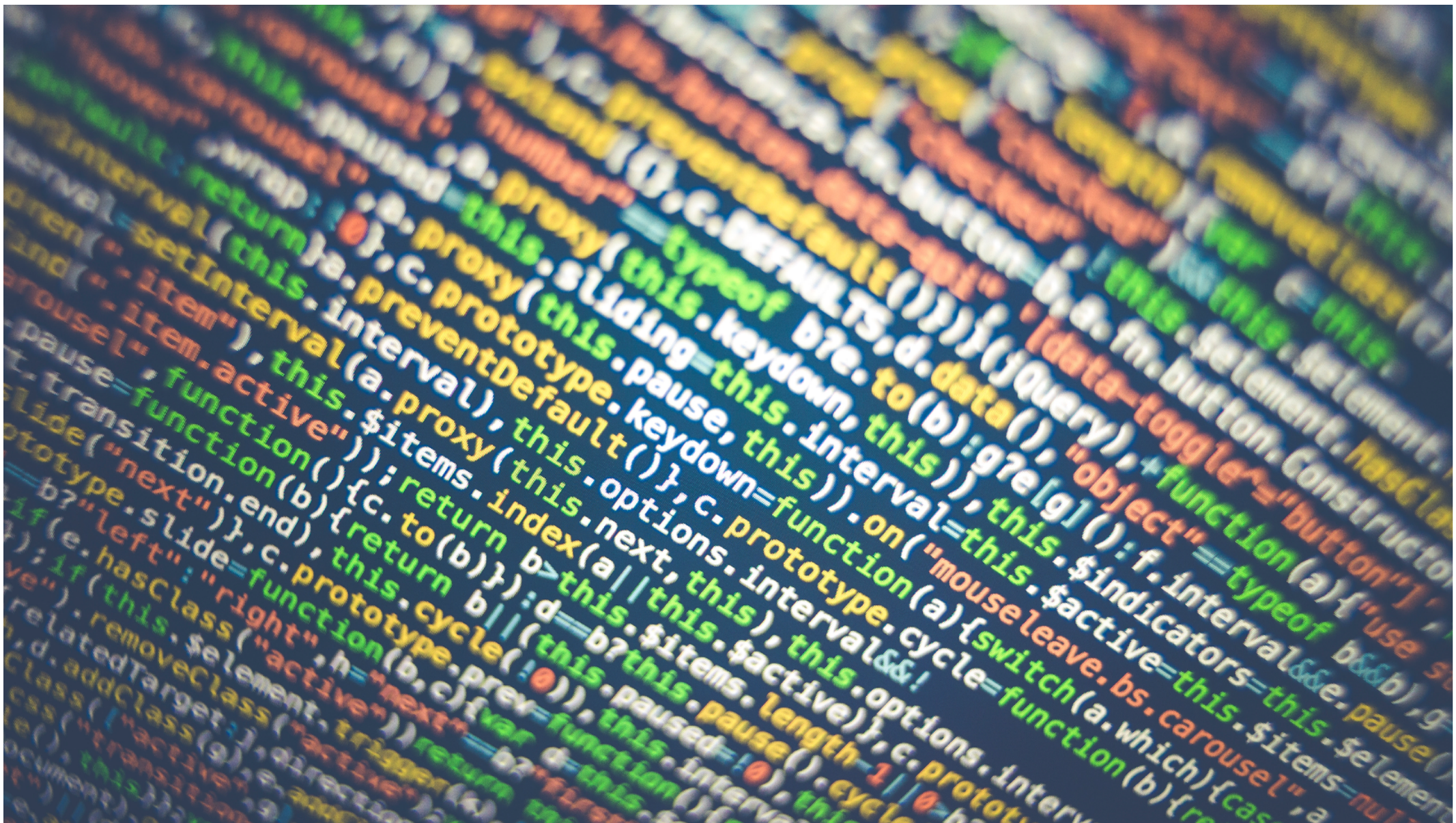
3. Ethical Issues

The use of deep learning algorithms, which feed off data for the purposes of personalization and assistance with decision-making, has given rise to the fear that social inequalities are being embedded in decision algorithms. In fact, much of the recent controversy surrounding this issue concerns discrimination towards certain minorities or based on gender (particularly black people, women and people living in deprived areas). American experience has also brought us several similar examples of the effects of discrimination in the field of crime prevention.

Because systems that incorporate AI technology are invading our daily lives, we legitimately expect them to act in accordance with our laws and social standards

Because systems that incorporate AI technology are invading our daily lives, we legitimately expect them to act in accordance with our laws and social standards. It is therefore essential that legislation and ethics control the performance of AI systems. Since we are currently unable to guarantee *a priori* the performance of a machine learning system (the formal certification of machine learning is still currently a subject of research), compliance with this requirement necessitates the development of procedures, tools and methods which will allow us to audit these systems in order to evaluate their conformity to our legal and ethical frameworks. This is also vital in case of litigation between different parties who are objecting to decisions taken by AI systems.

Should data have an expiration date?



Other AI Challenges

1. Open AI

MOA

- {M}assive {O}nline {A}nalysis is a framework for online learning from data streams.

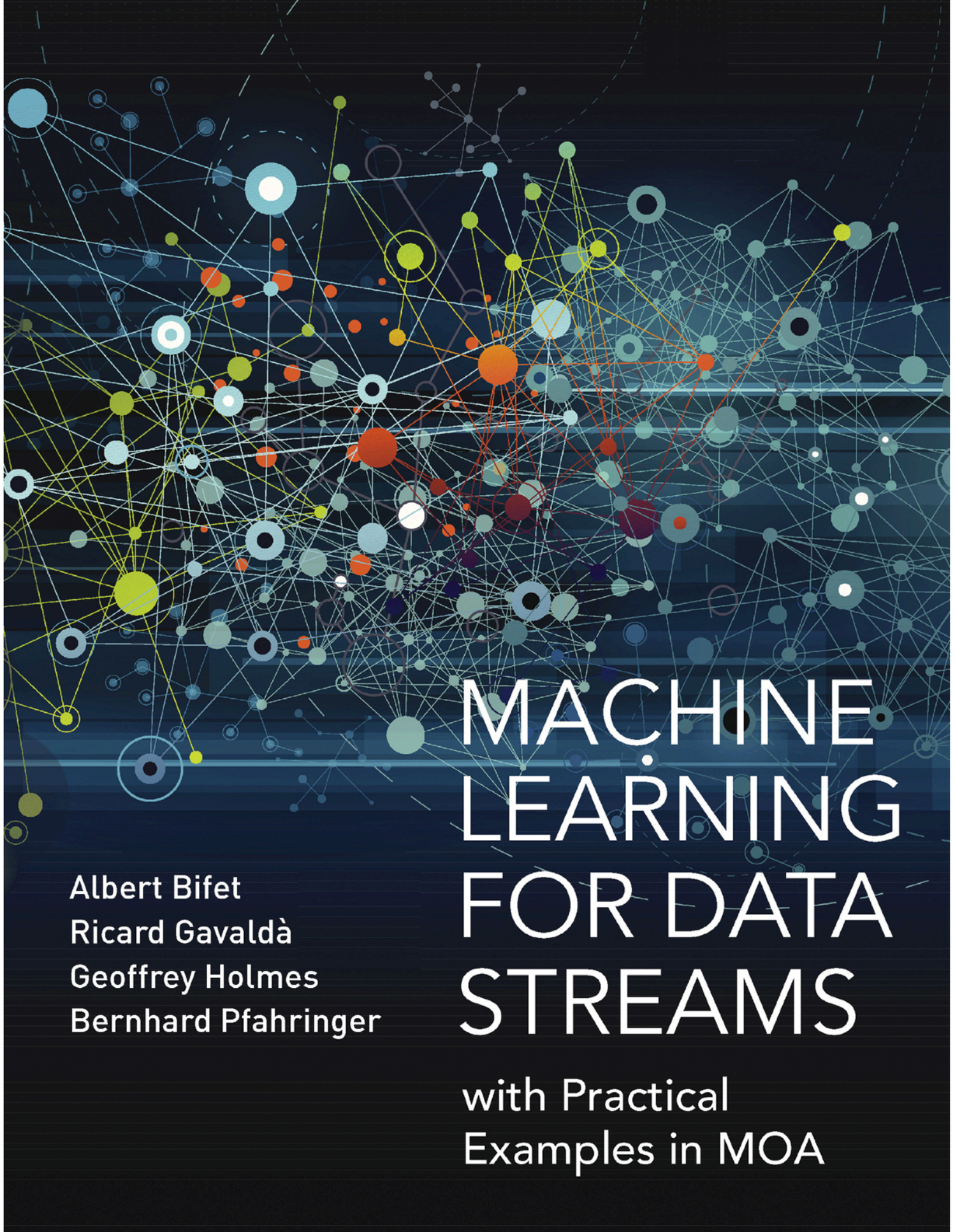
- It is closely related to WEKA

- It includes a collection of offline and online as well as tools for evaluation:

  - classification, regression

  - clustering, frequent pattern mining

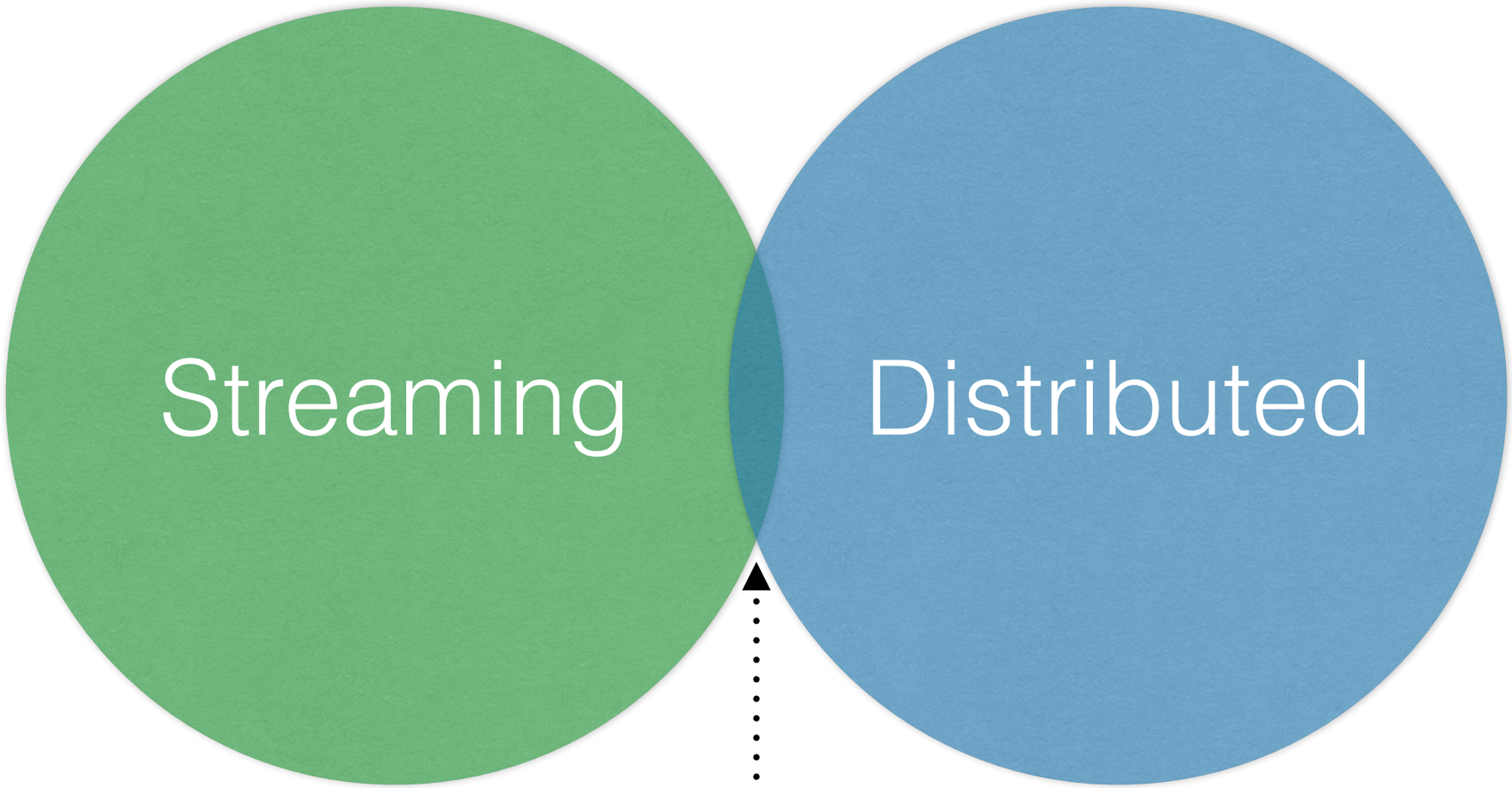
- Easy to extend, design and run experiments



The MIT Press

2. Distributed Data Stream Mining

Vision

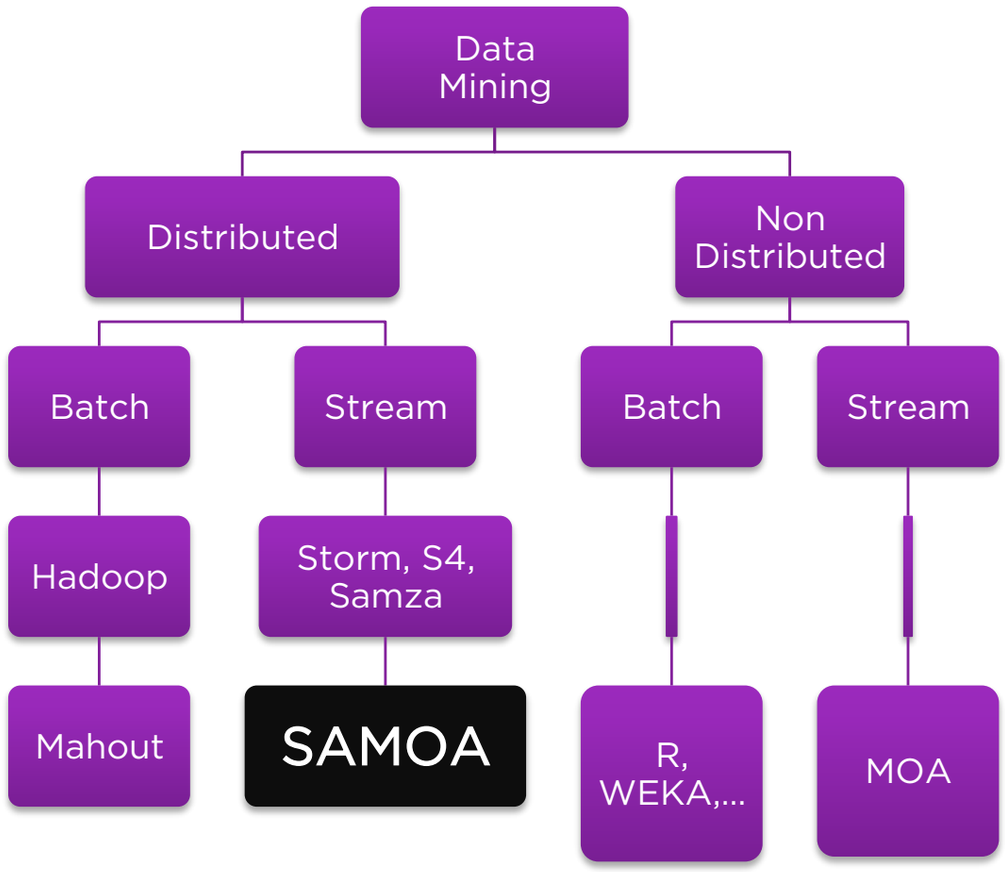


IoT Big Data Stream Mining

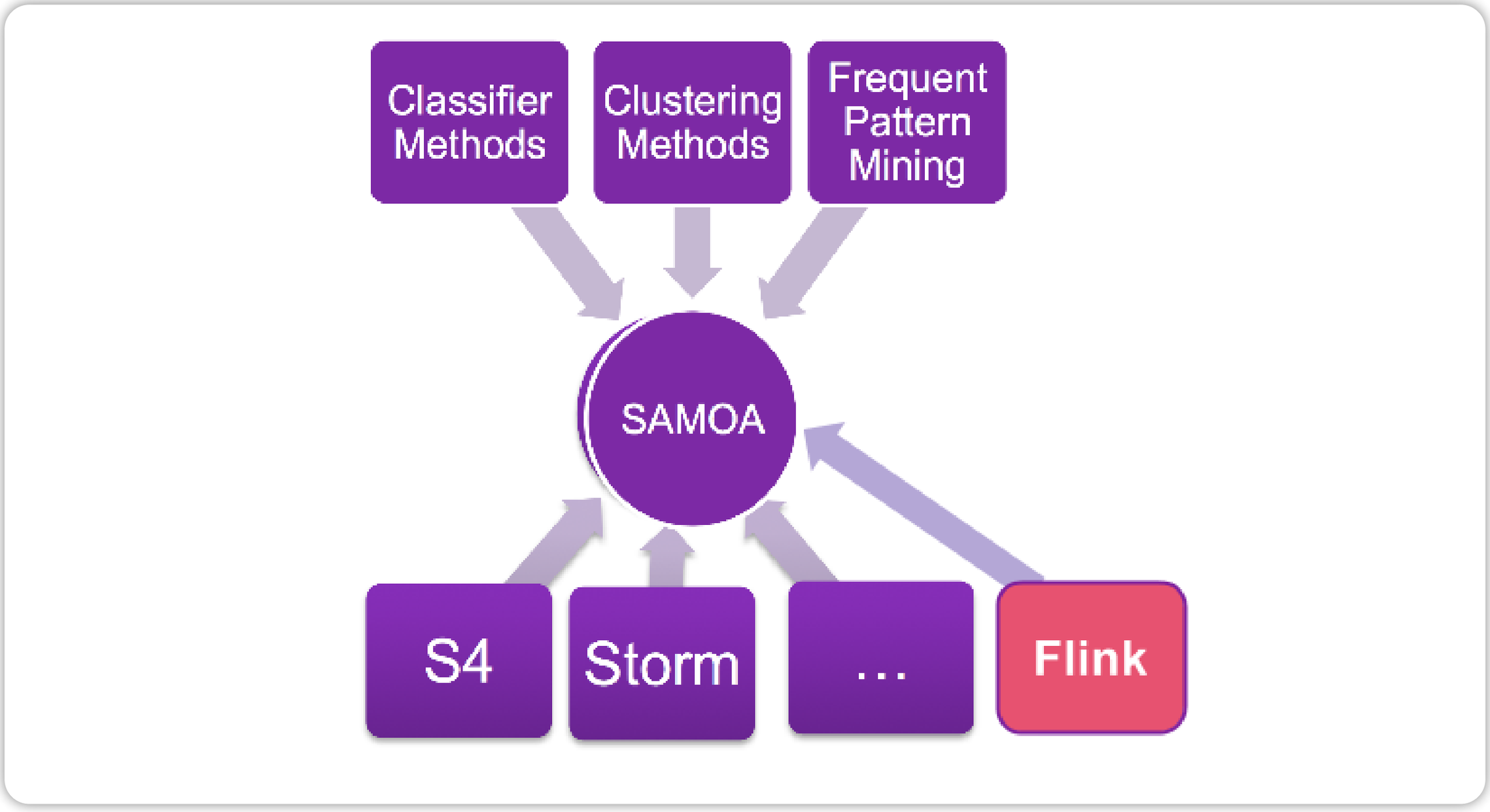
http://samoa-project.net

APACHE SAMOA

G. De Francisci Morales, A. Bifet: “SAMOA: Scalable Advanced Massive Online Analysis”. JMLR (2014)

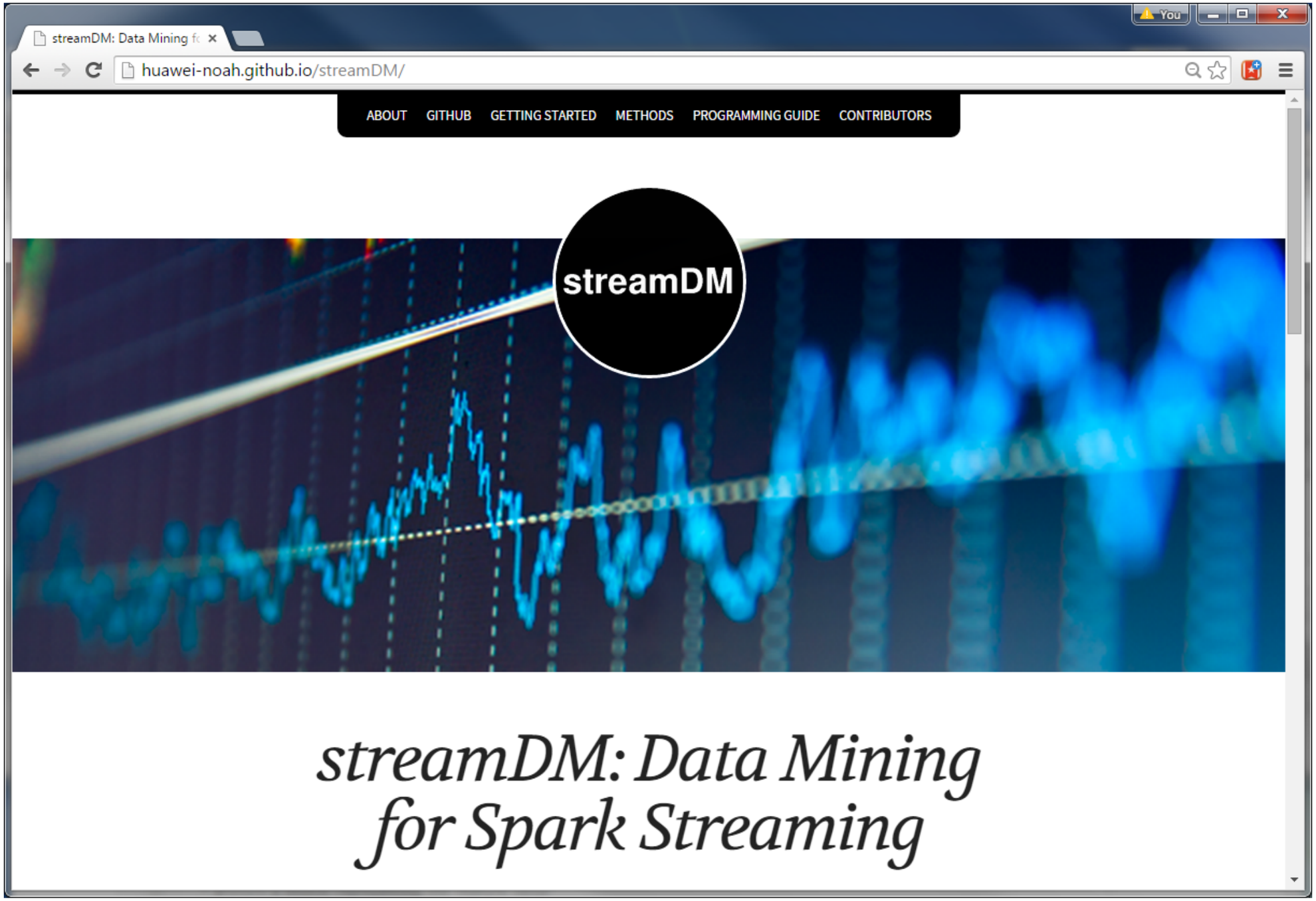


SAMOA ARCHITECTURE



http://huawei-noah/github.io/streamDM

StreamDM



scikit-multiflow

A multi-output/multi-label and stream data framework. Inspired by MOA and MEKA, following scikit-learn philosophy.

Github Repository

Photo credit: freepik

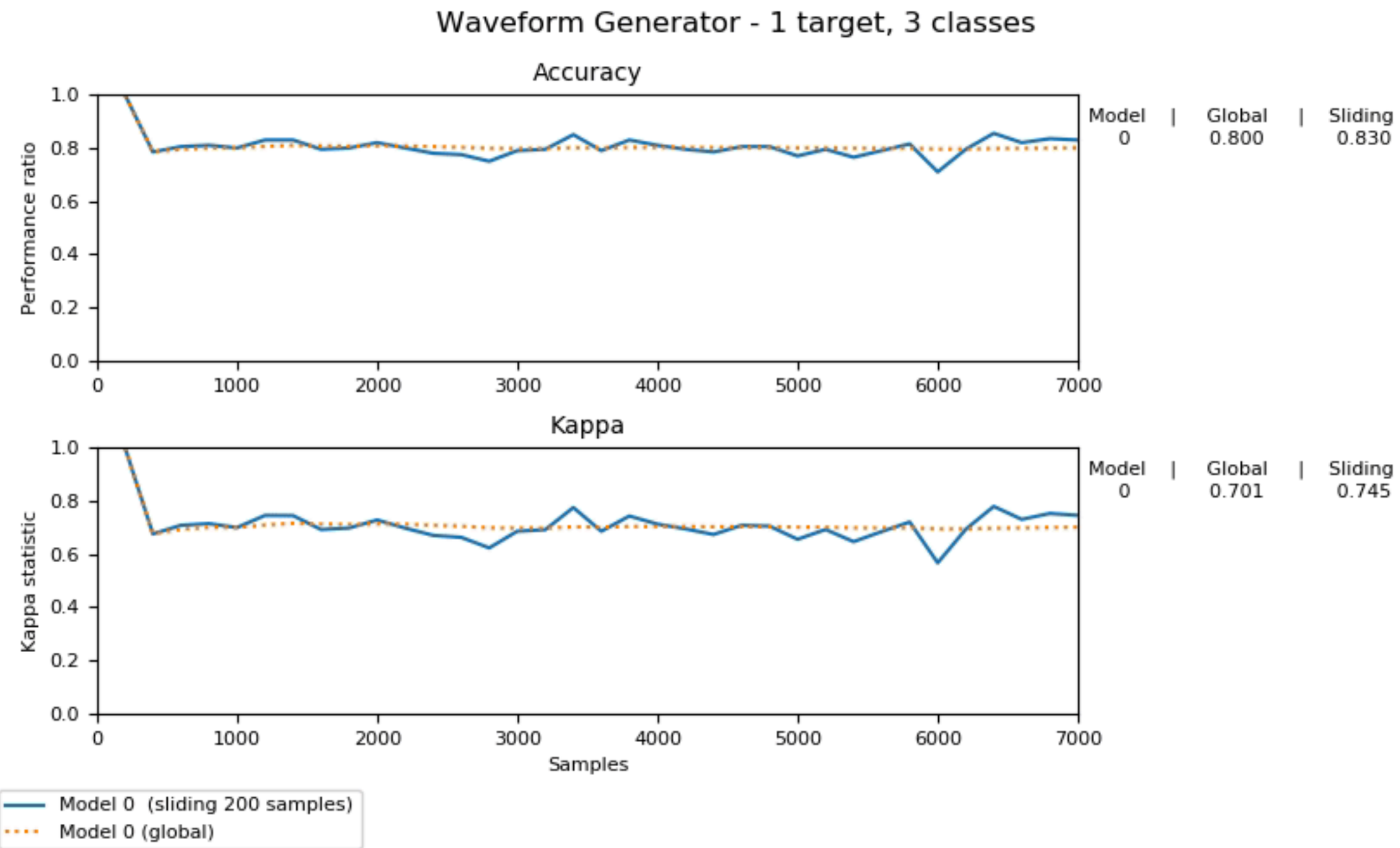
Mainly supported by: TELECOM ParisTech, ÉCOLE POLYTECHNIQUE UNIVERSITÉ PARIS-SACLAY

GITHUB

scikit-multiflow

```
from skmultiflow.data.generators.waveform_generator import Wave
from skmultiflow.classification.trees.hoeffding_tree import Hoe
from skmultiflow.evaluation.evaluate_prequential import Evaluat

# 1. Create a stream
stream = WaveformGenerator()
stream.prepare_for_use()

# 2. Instantiate the HoeffdingTree classifier
ht = HoeffdingTree()

# 3. Setup the evaluator
eval = EvaluatePrequential(show_plot=True, pretrain_size=1000,

# 4. Run evaluation
eval.eval(stream=stream, classifier=ht)
```

scikit-multiflow



scikit-multiflow

Jesse Read
Ecole Polytechnique
France

Jacob Montiel
Telecom ParisTech
France

Summary

- Machine Learning for Data Streams useful for finding approximate solutions with reasonable amount of time & limited resources

- Challenges:
  - Open AI
  - Green AI
  - Explainable AI
  - Ethical Issues
  - Distributed Data Stream Mining

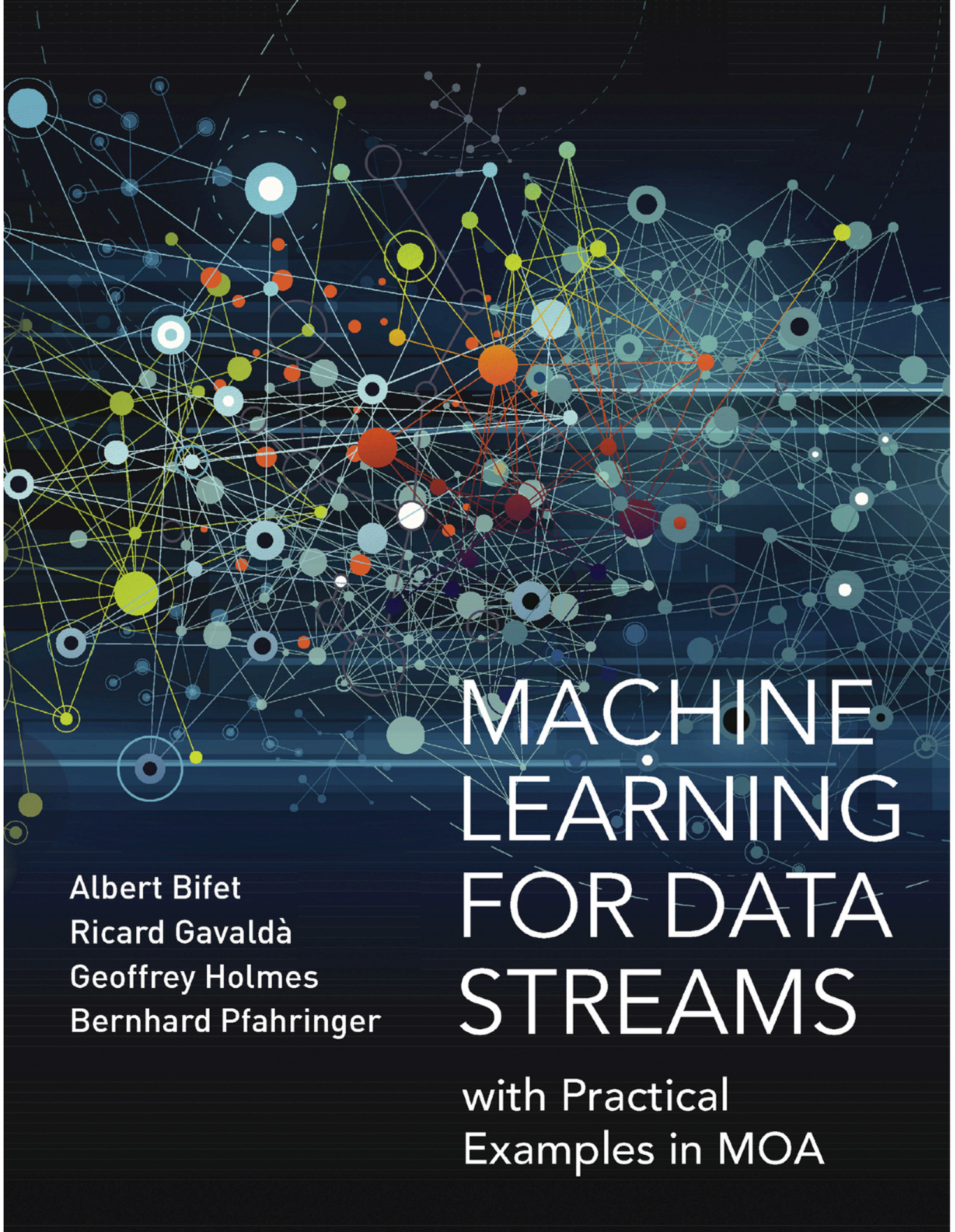
Summary

- Green AI
- Explainable AI
- Ethical Issues
- Open AI
- Distributed Data Stream Mining

Green Data Mining

1st International Workshop on Energy Efficient Data Mining and Knowledge Discovery

Co-located with ECML PKDD 2018



The MIT Press

Thanks!

@abifet

Machine Learning for Data Streams

Albert Bifet (@abifet)

DATAIA-JST International
Symposium on Data Science and AI
11 July 2018