Machine Learning for Data Streams

Albert Bifet (@abifet)

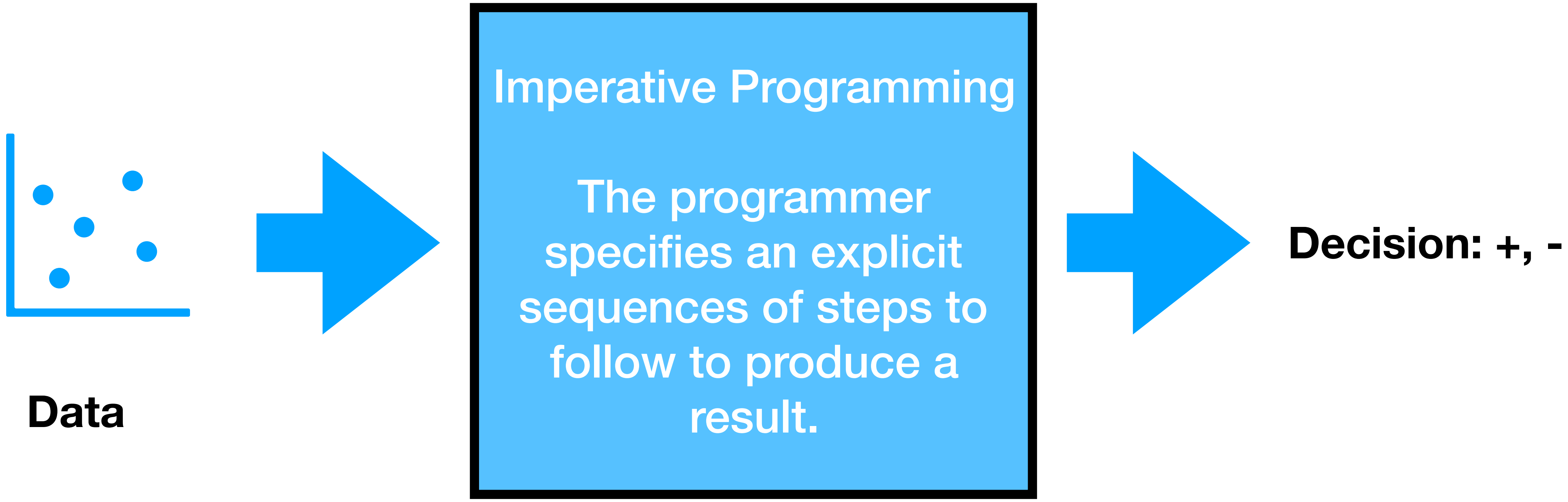
Cisco-Ecole Polytechnique Symposium 2018,
10 April 2018

Machine Learning

- **Machine learning** is a type of artificial intelligence (**AI**) that provides computers with the ability to learn without being explicitly programmed.

- **Machine learning** focuses on the development of computer programs **that can teach themselves to grow** and change when exposed to new data.

Machine Learning



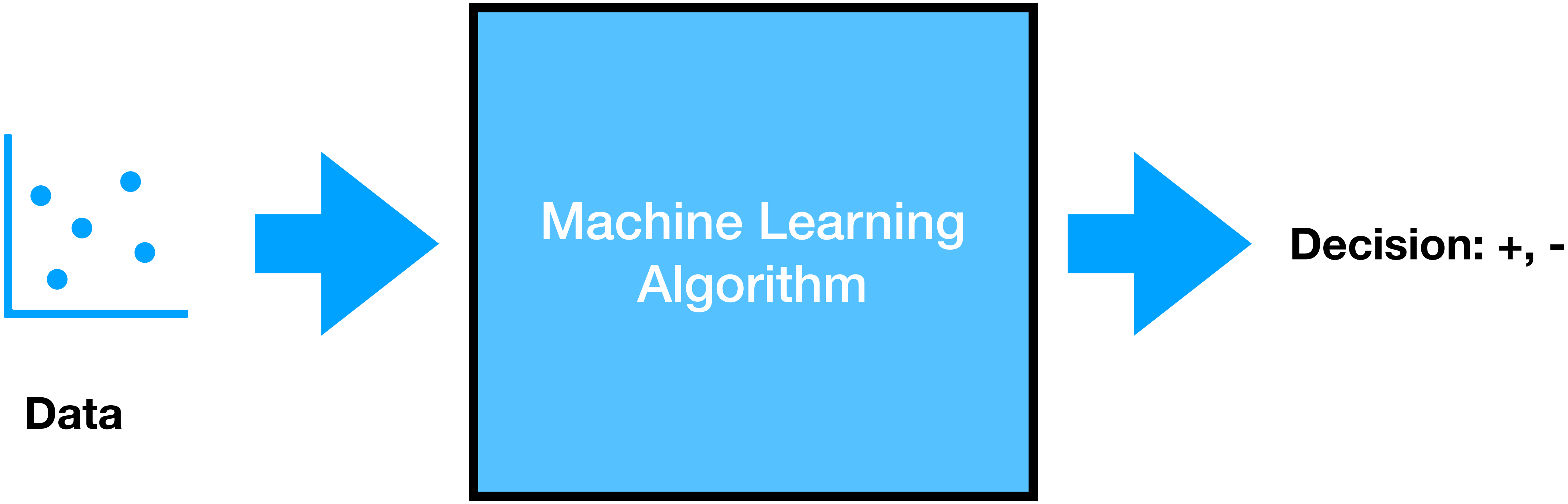
Data

Imperative Programming

The programmer specifies an explicit sequences of steps to follow to produce a result.

Decision: +, -

Machine Learning



Data

Machine Learning Algorithm

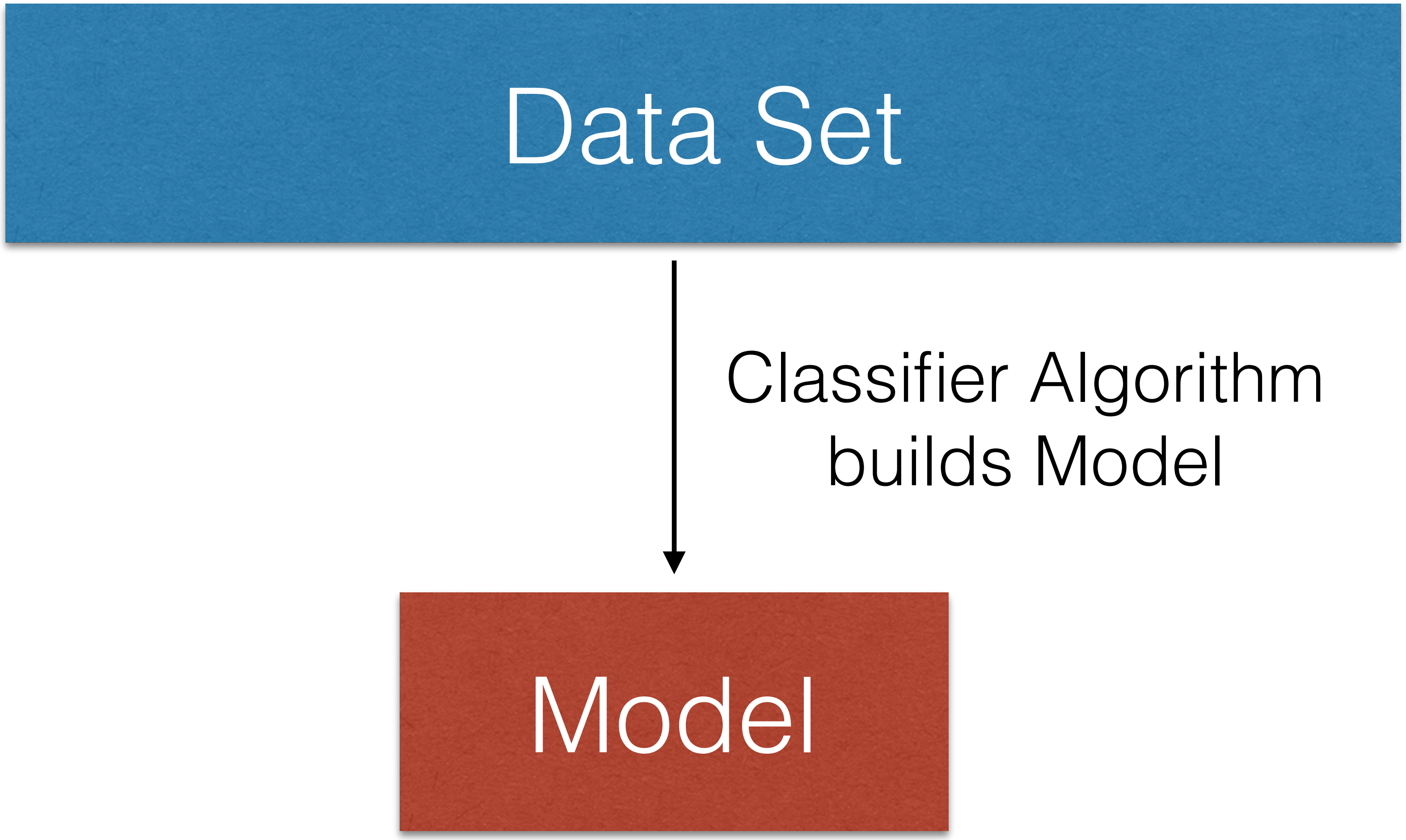
Decision: +, -

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



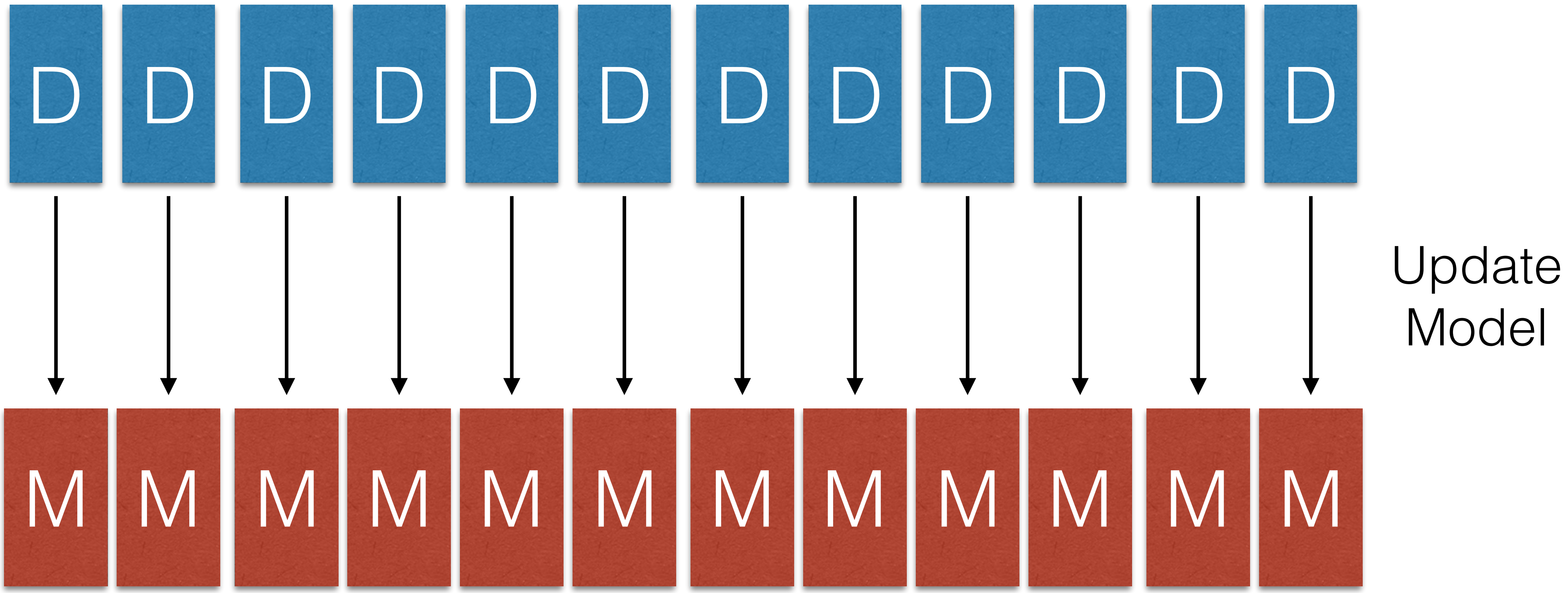
Data Set

Classifier Algorithm
builds Model

Model

Analytic Standard Approach

Finite training sets
Static models

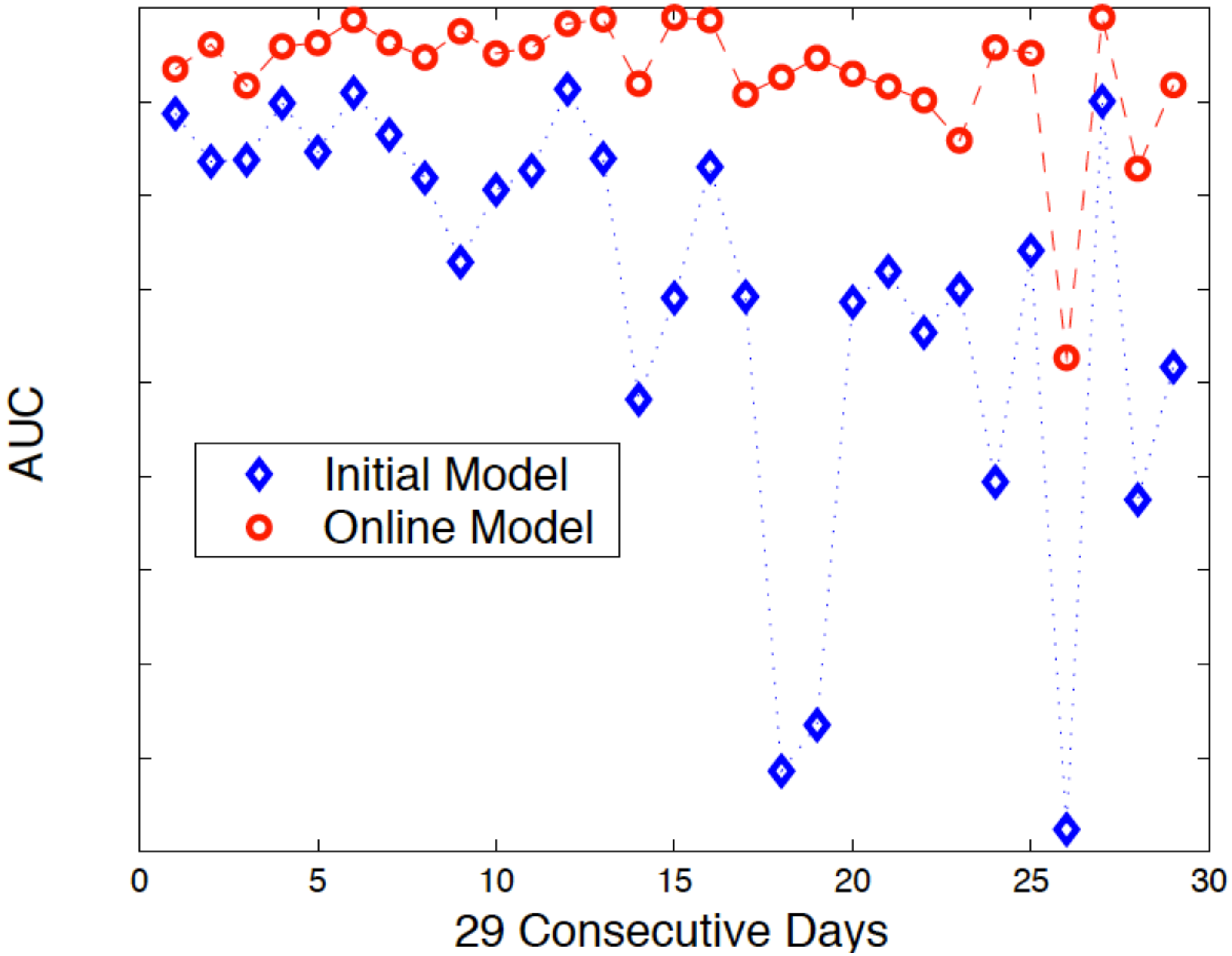


Data Stream Approach

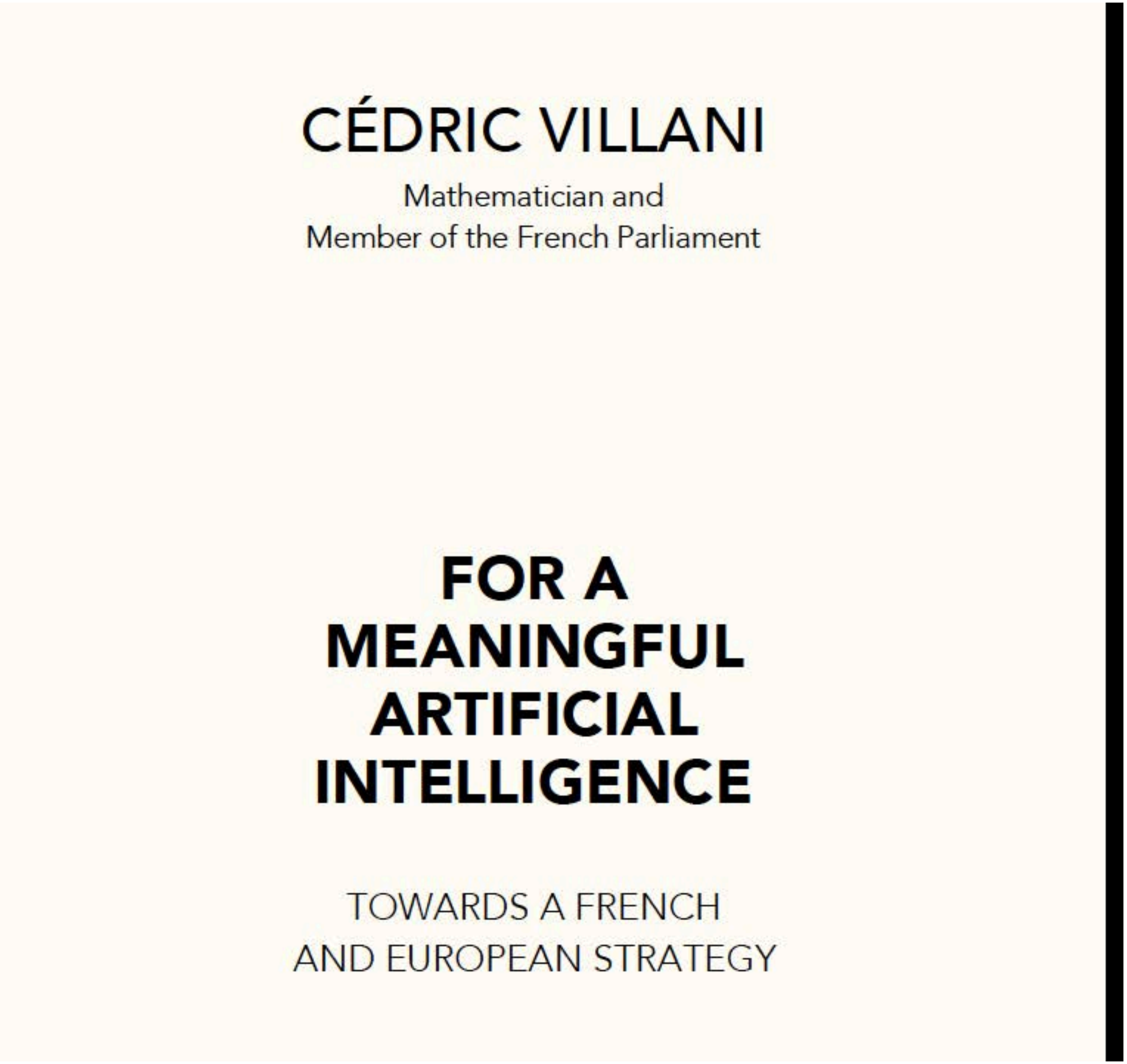
Infinite training sets
Dynamic models

Adversarial Learning

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



AI Challenges



CÉDRIC VILLANI

Mathematician and
Member of the French Parliament

FOR A MEANINGFUL ARTIFICIAL INTELLIGENCE

TOWARDS A FRENCH
AND EUROPEAN STRATEGY

Cédric Villani and Marc Shoenauer

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.

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

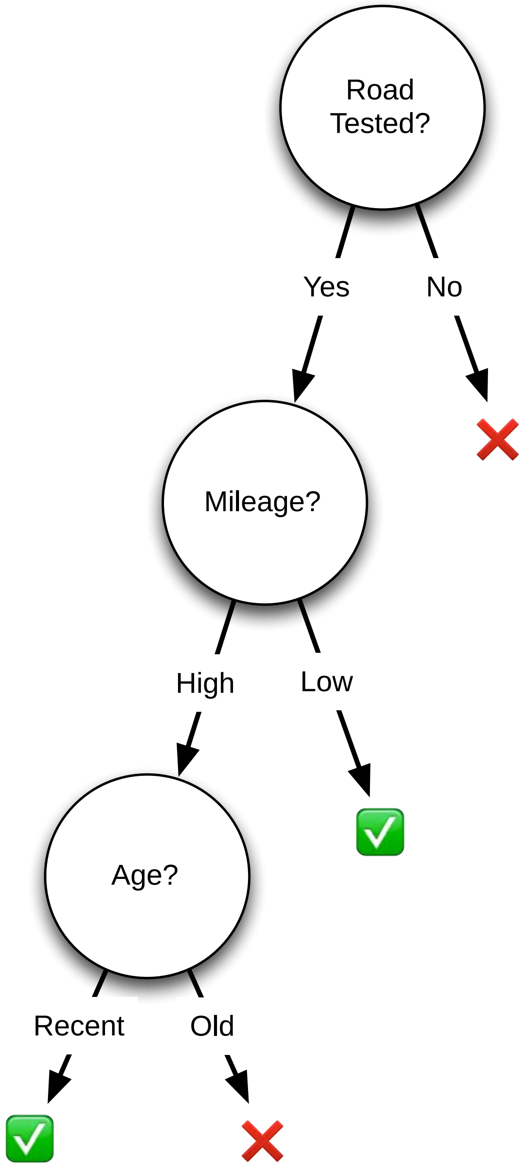
Finally, our digital society could not be governed by black box algorithms: artificial intelligence is going to play a decisive role in critical domains for human flourishing (health, banking, housing, etc) and there is currently a high risk of embedding existing discrimination into AI algorithms or 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.

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

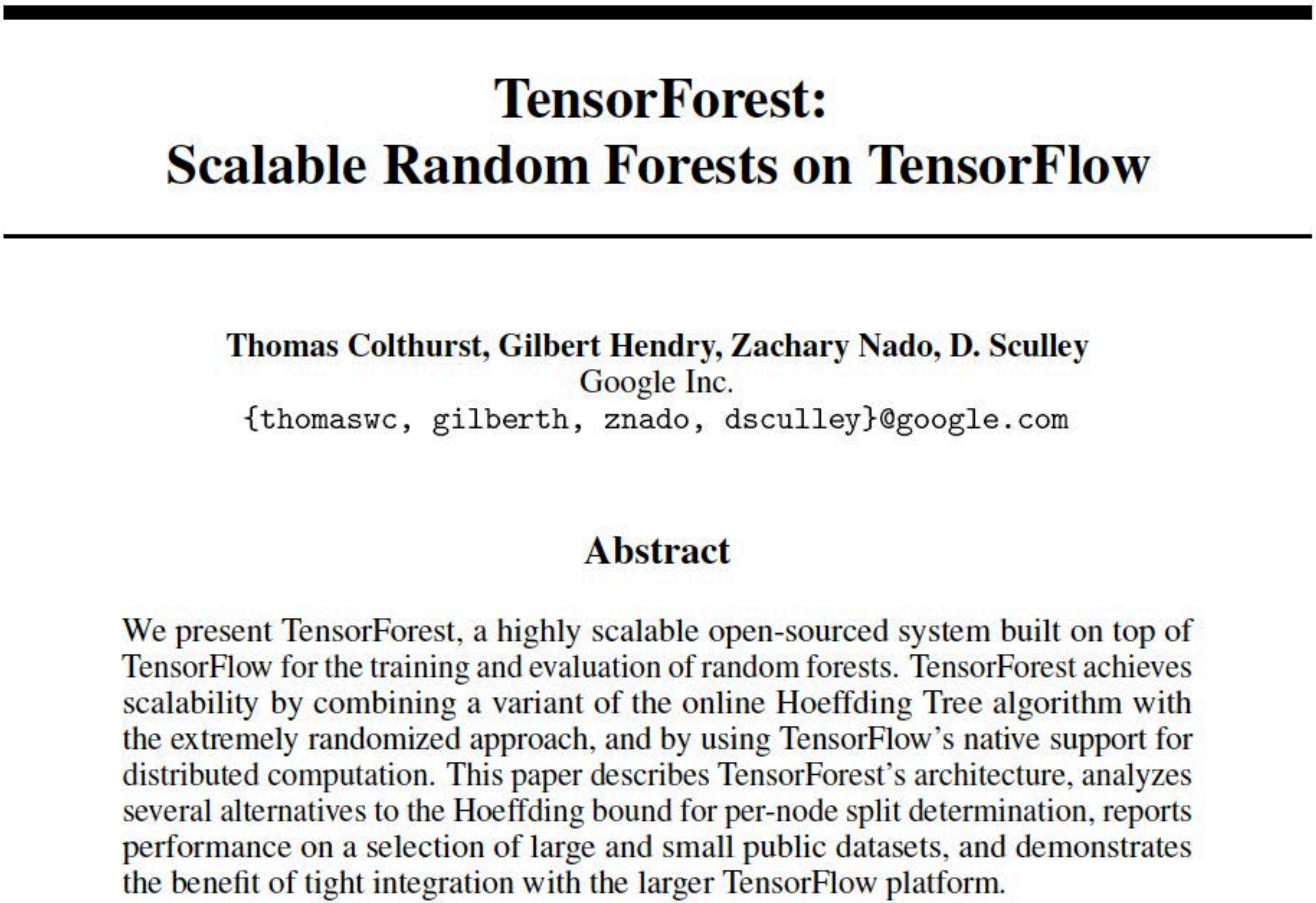
Car deal?



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

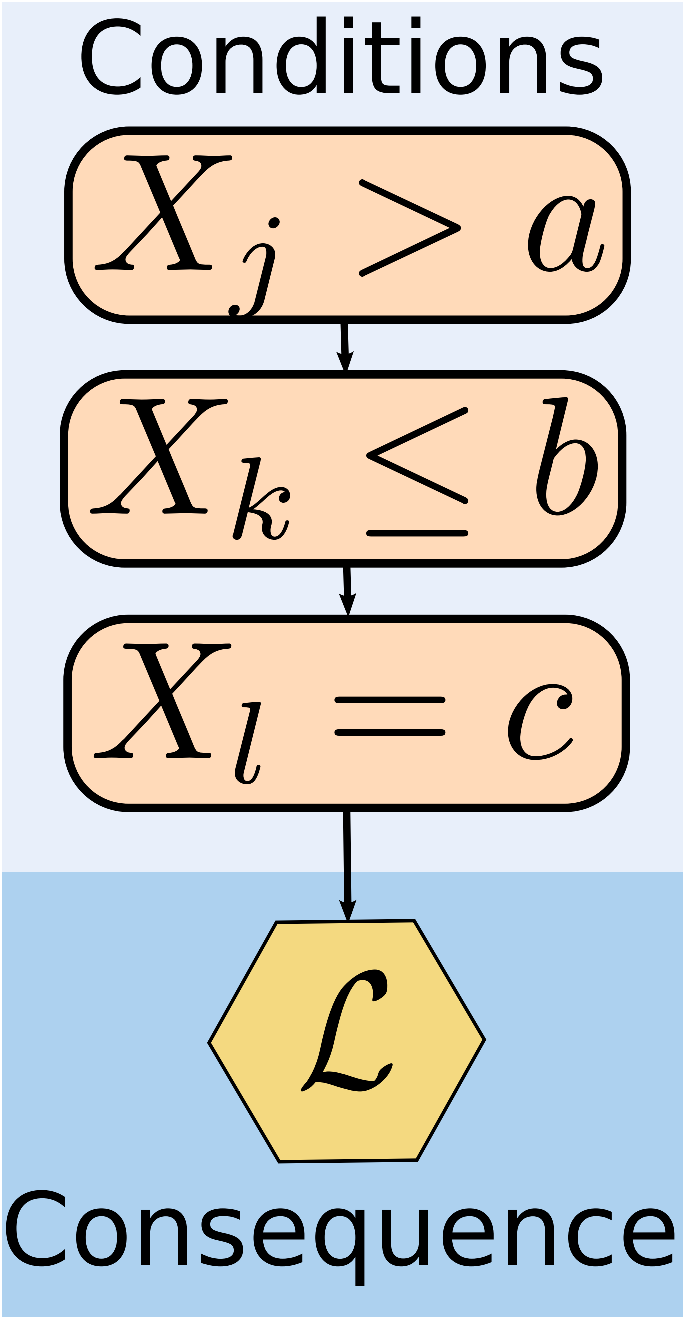
Thomas Colthurst, Gilbert Hendry, Zachary Nado, D. Sculley
Google Inc.
{thomaswc, gilberth, znado, dsculley}@google.com

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.

Rules

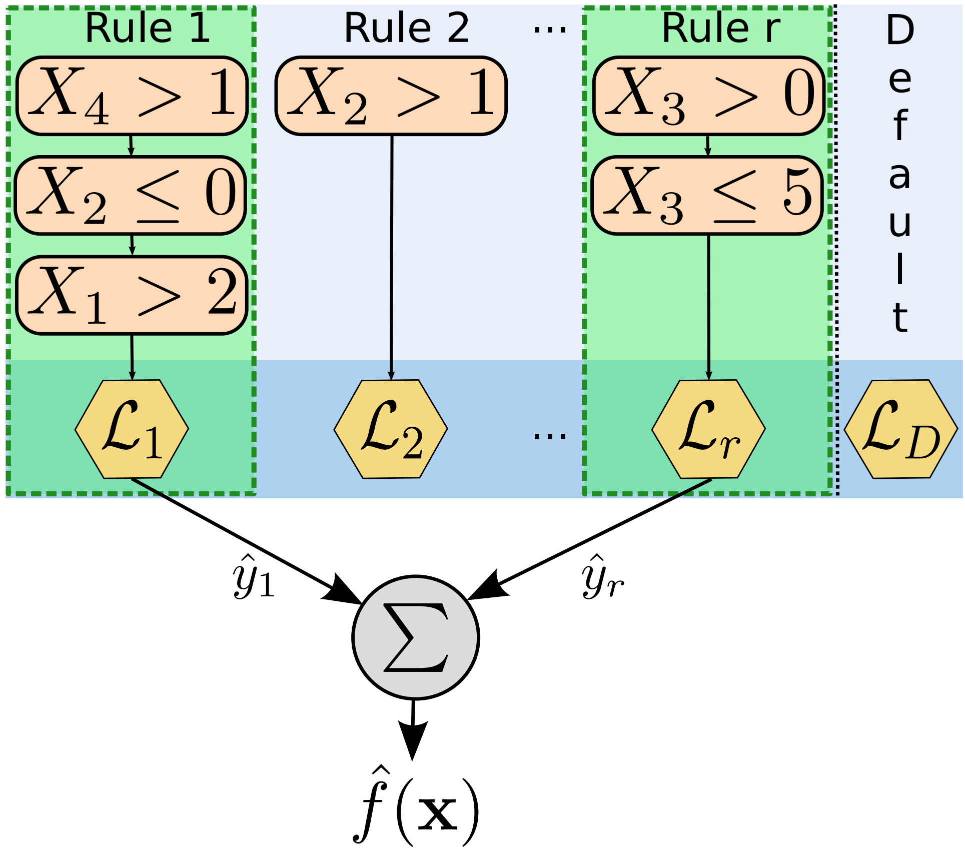
- Problem: very large decision trees have context that is complex and hard to understand
- Rules: self-contained, modular, easier to interpret, no need to cover universe
- $\mathcal{L}$ keeps sufficient statistics to:
  - make predictions
  - expand the rule
  - detect changes and anomalies



Adaptive Model Rules

E. Almeida, C. Ferreira, J. Gama. "Adaptive Model Rules from Data Streams." ECML-PKDD ‘13

- Ruleset: ensemble of rules
- Rule prediction: mean, linear model
- Ruleset prediction
  - Weighted avg. of predictions of rules covering instance x
  - Weights inversely proportional to error
  - Default rule covers uncovered instances



E.g: $\mathbf{x} = [4, -1, 1, 2]$

$$\hat{f}(\mathbf{x}) = \sum_{R_l \in S(\mathbf{x}_i)} \theta_l \hat{y}_l,$$

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

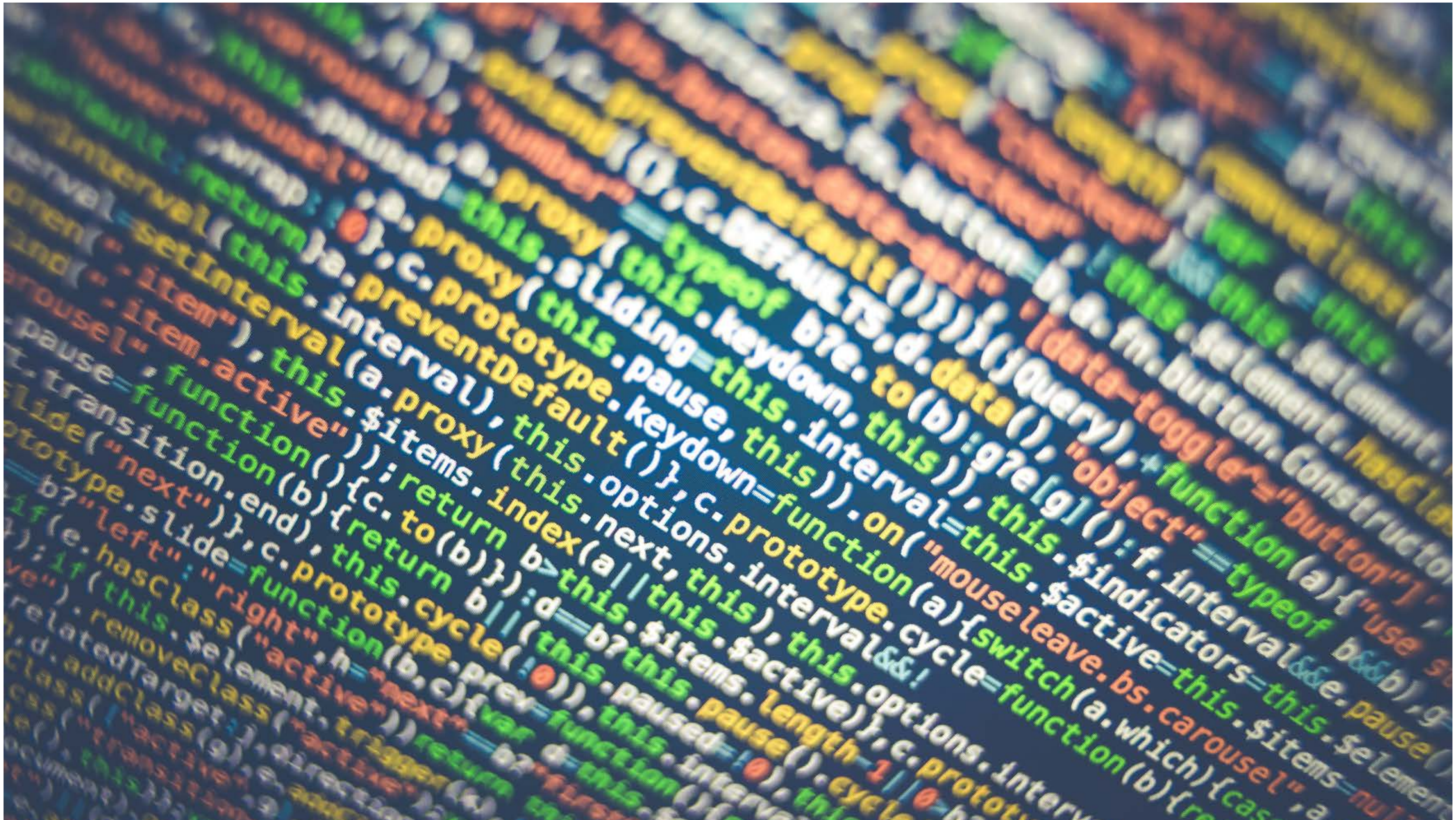
3. Ethical Issues

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. 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.

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

Should data have an expiration date?



Other AI Challenges

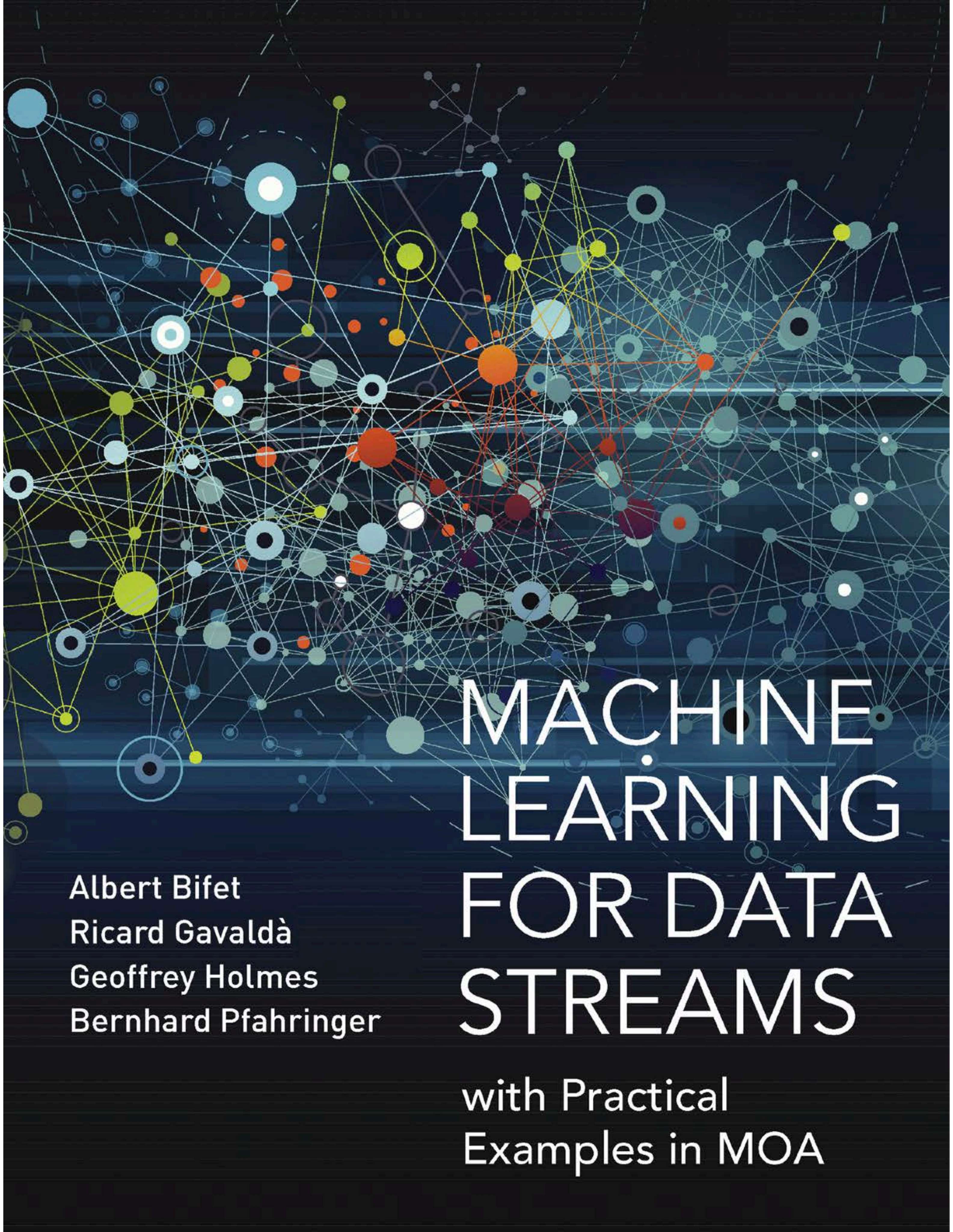
1. Open AI

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



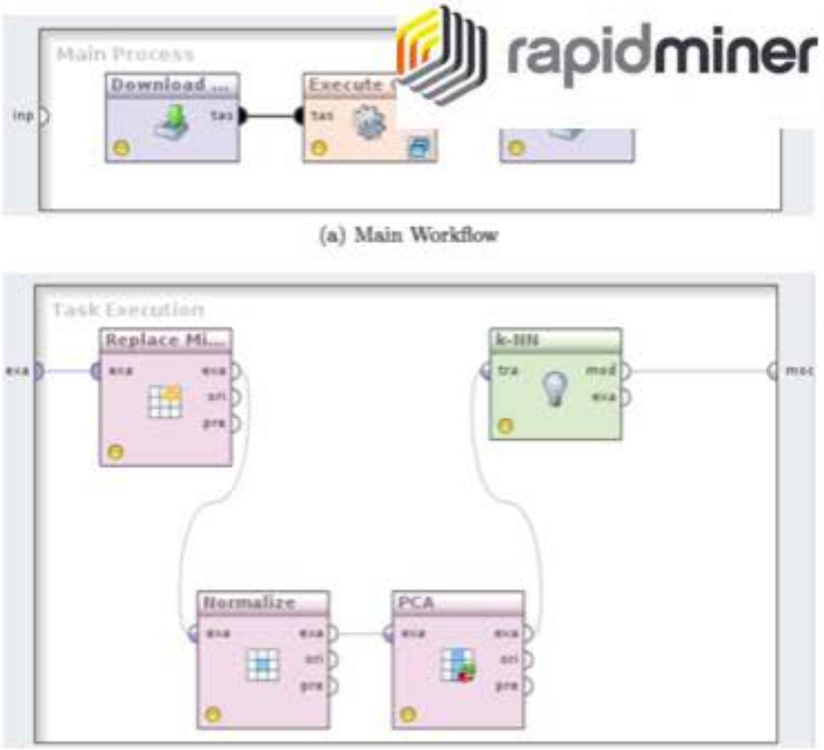
Albert Bifet
Ricard Gavaldà
Geoffrey Holmes
Bernhard Pfahringer

MACHINE LEARNING FOR DATA STREAMS

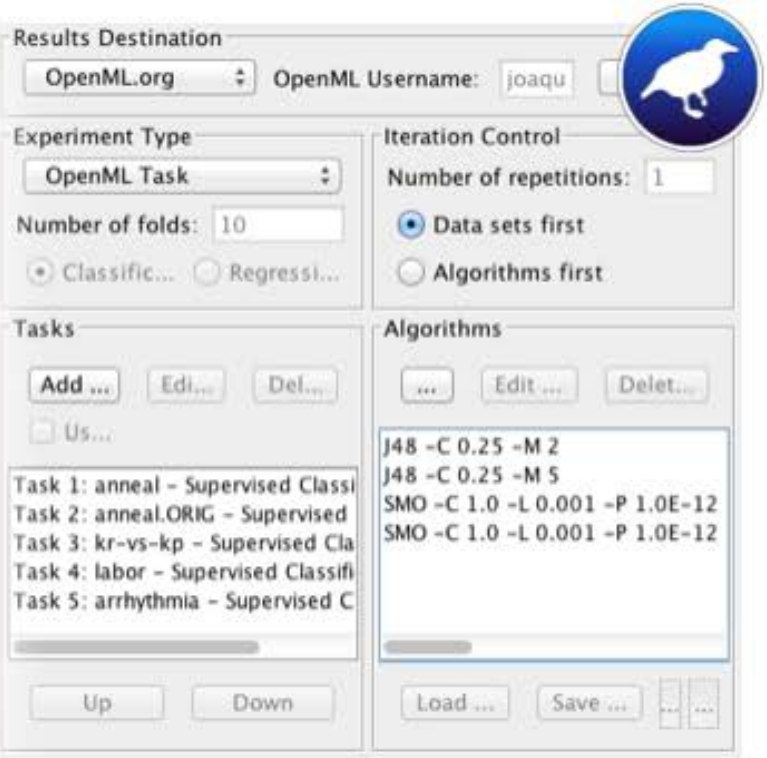
with Practical Examples in MOA

OpenML

NETWORKED SCIENCE AND MACHINE LEARNING



(a) Main Workflow


Results Destination

OpenML.org — OpenML Username: joaqu

Experiment Type

OpenML Task

Number of folds: 10

○ Classific... ○ Regressi...

Iteration Control

Number of repetitions: 1

● Data sets first

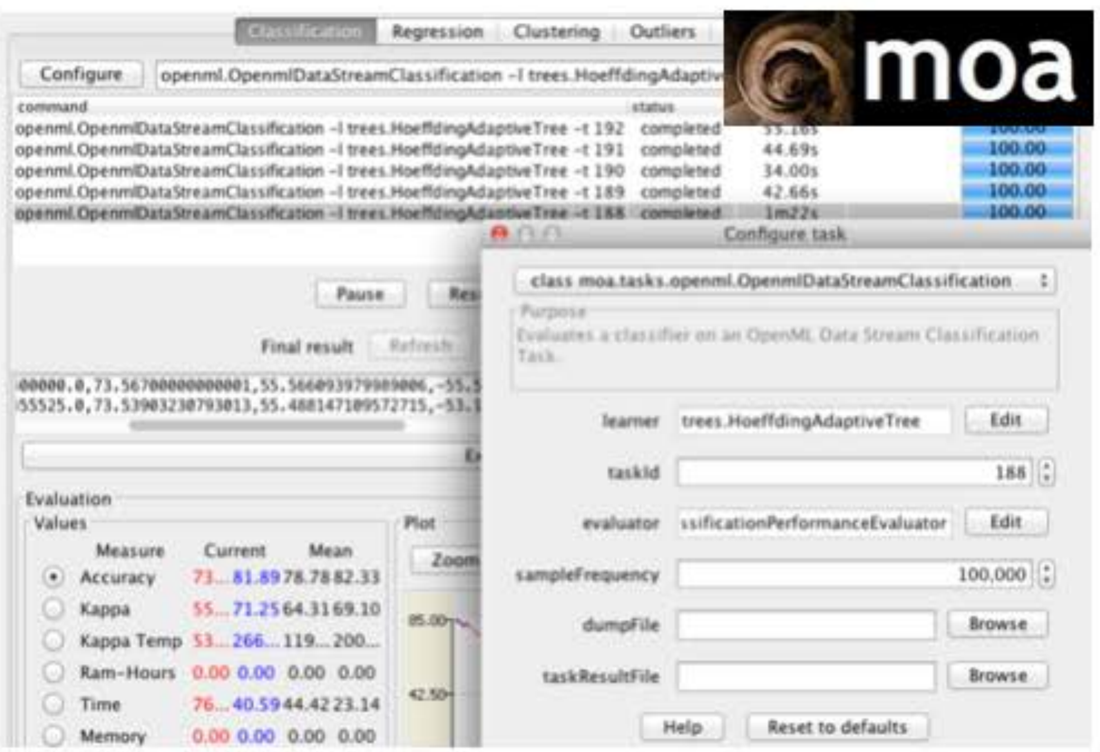
○ Algorithms first

Tasks

Task 1: anneal - Supervised Classi
Task 2: anneal.ORIG - Supervised
Task 3: kr-vs-kp - Supervised Cla
Task 4: labor - Supervised Classifi
Task 5: arrhythmia - Supervised C

Algorithms

```
J48 -C 0.25 -M 2
J48 -C 0.25 -M 5
SMO -C 1.0 -L 0.001 -P 1.0E-12
SMO -C 1.0 -L 0.001 -P 1.0E-12
```



Configure: openml.OpenmlDataStreamClassification -l trees.HoeffdingAdaptiv

```
command                                                                  status
openml.OpenmlDataStreamClassification -l trees.HoeffdingAdaptiveTree -t 192  completed   55.16s   100.00
openml.OpenmlDataStreamClassification -l trees.HoeffdingAdaptiveTree -t 191  completed   44.69s   100.00
openml.OpenmlDataStreamClassification -l trees.HoeffdingAdaptiveTree -t 190  completed   34.00s   100.00
openml.OpenmlDataStreamClassification -l trees.HoeffdingAdaptiveTree -t 189  completed   42.66s   100.00
openml.OpenmlDataStreamClassification -l trees.HoeffdingAdaptiveTree -t 188  completed   1m22s    100.00
```

```
00000,0,73.56700000000001,55.566093979989006,-55.5
55525,0,73.53903230793013,55.488147109572715,-53.1
```

Evaluation

Values

	Measure	Current		Mean	
●	Accuracy	73...	81.89	78.78	82.33
○	Kappa	55...	71.25	64.31	69.10
○	Kappa Temp	53...	266...	119...	200...
○	Ram-Hours	0.00	0.00	0.00	0.00
○	Time	76...	40.59	44.42	23.14
○	Memory	0.00	0.00	0.00	0.00

Plot: 85.00, 42.50

Configure task

class moa.tasks.openml.OpenmlDataStreamClassification

Purpose: Evaluates a classifier on an OpenML Data Stream Classification Task.

learner: trees.HoeffdingAdaptiveTree [Edit]

taskId: 188

evaluator: ssificationPerformanceEvaluator [Edit]

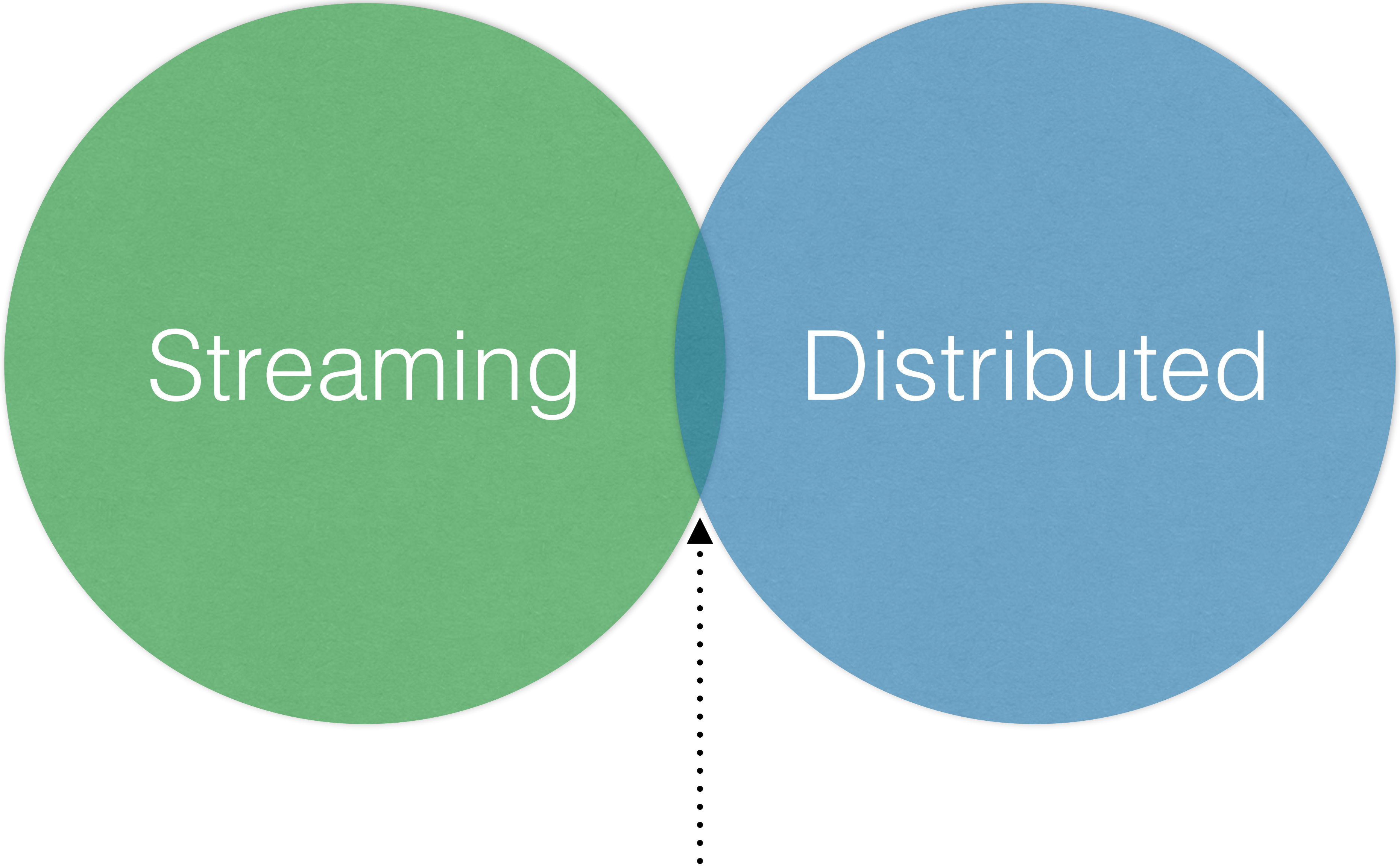
sampleFrequency: 100,000

dumpFile: [Browse]

taskResultFile: [Browse]

2. Distributed Data Stream Mining

Vision

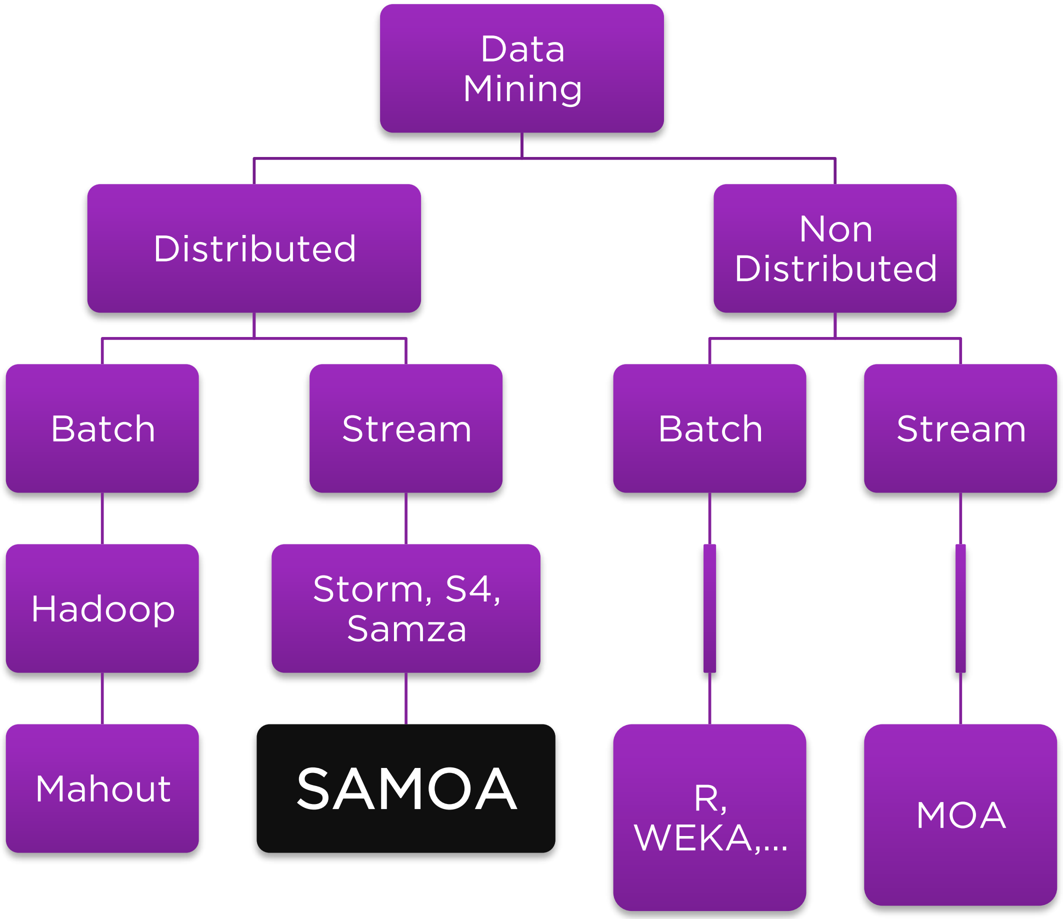


IoT Big Data Stream Mining

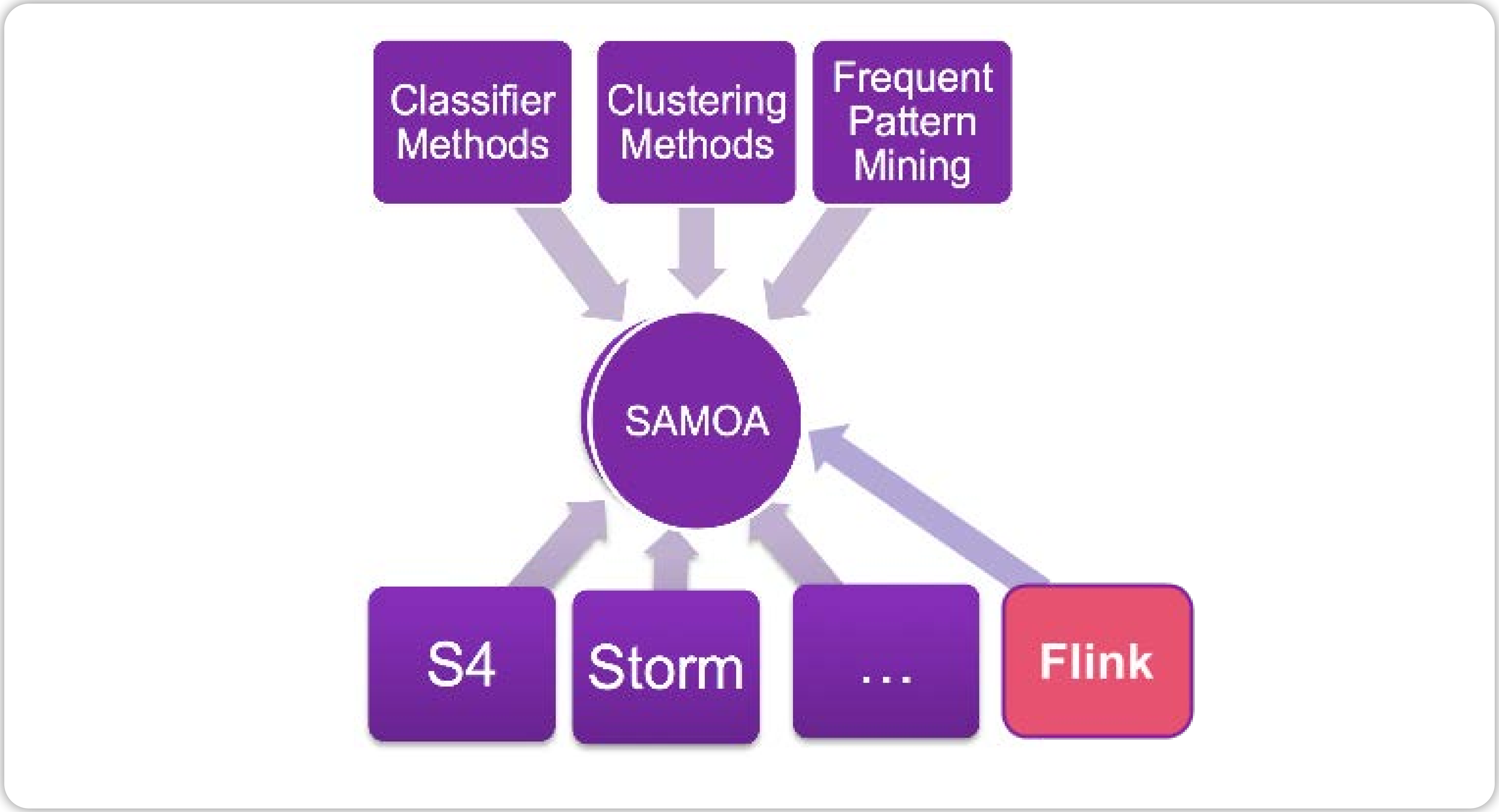
http://samoa-project.net

APACHE SAMOA

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

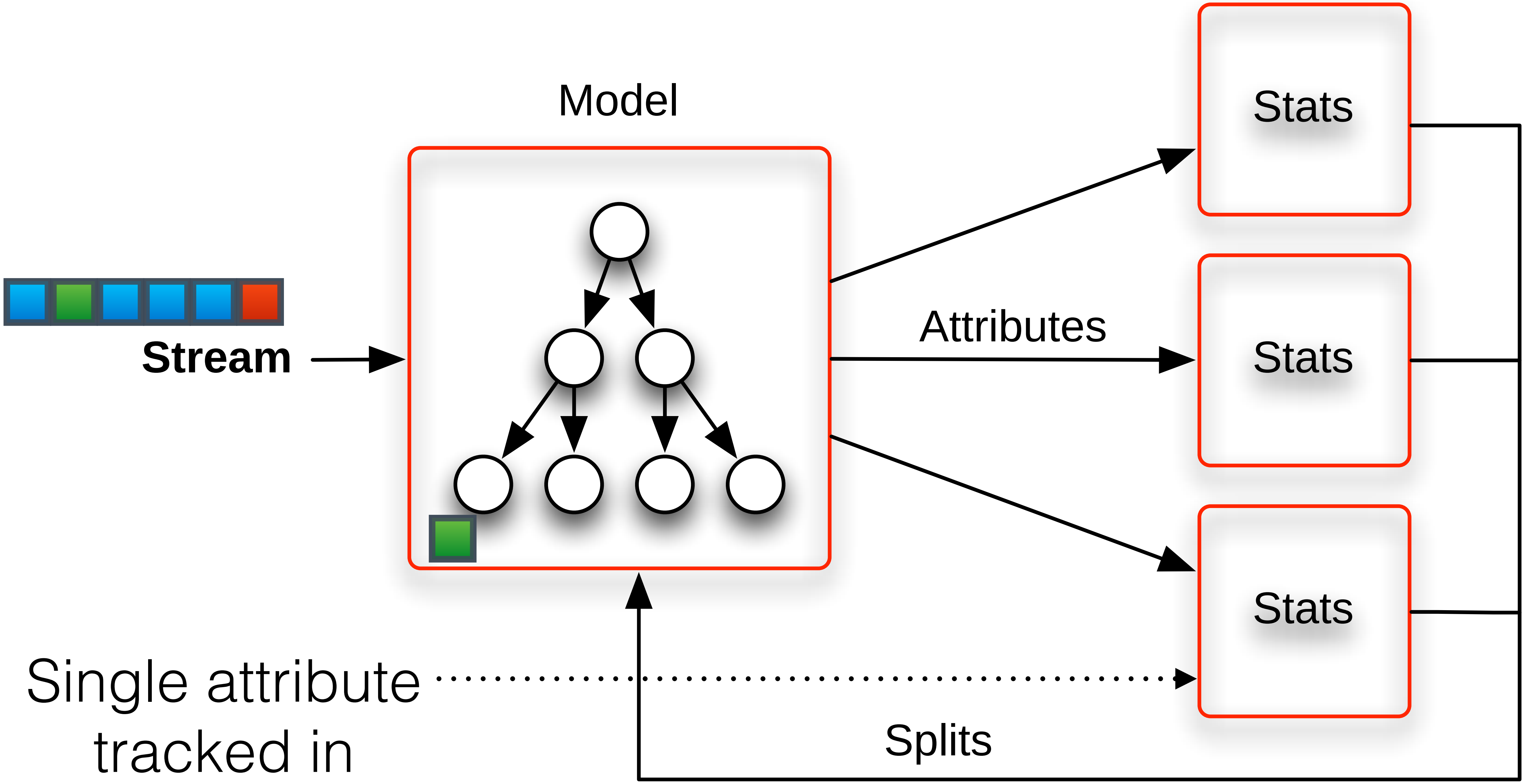


SAMOA ARCHITECTURE



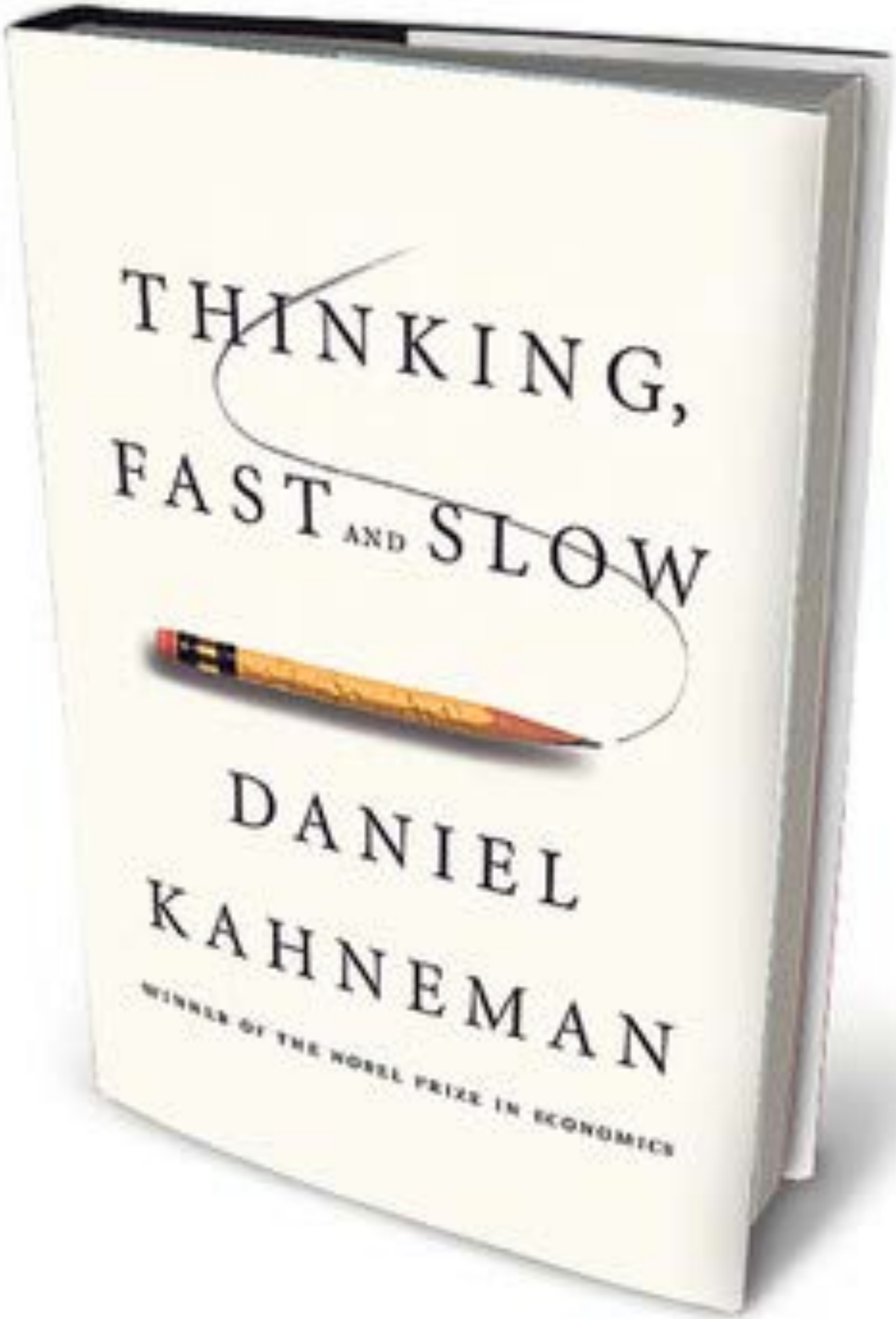
Vertical Partitioning

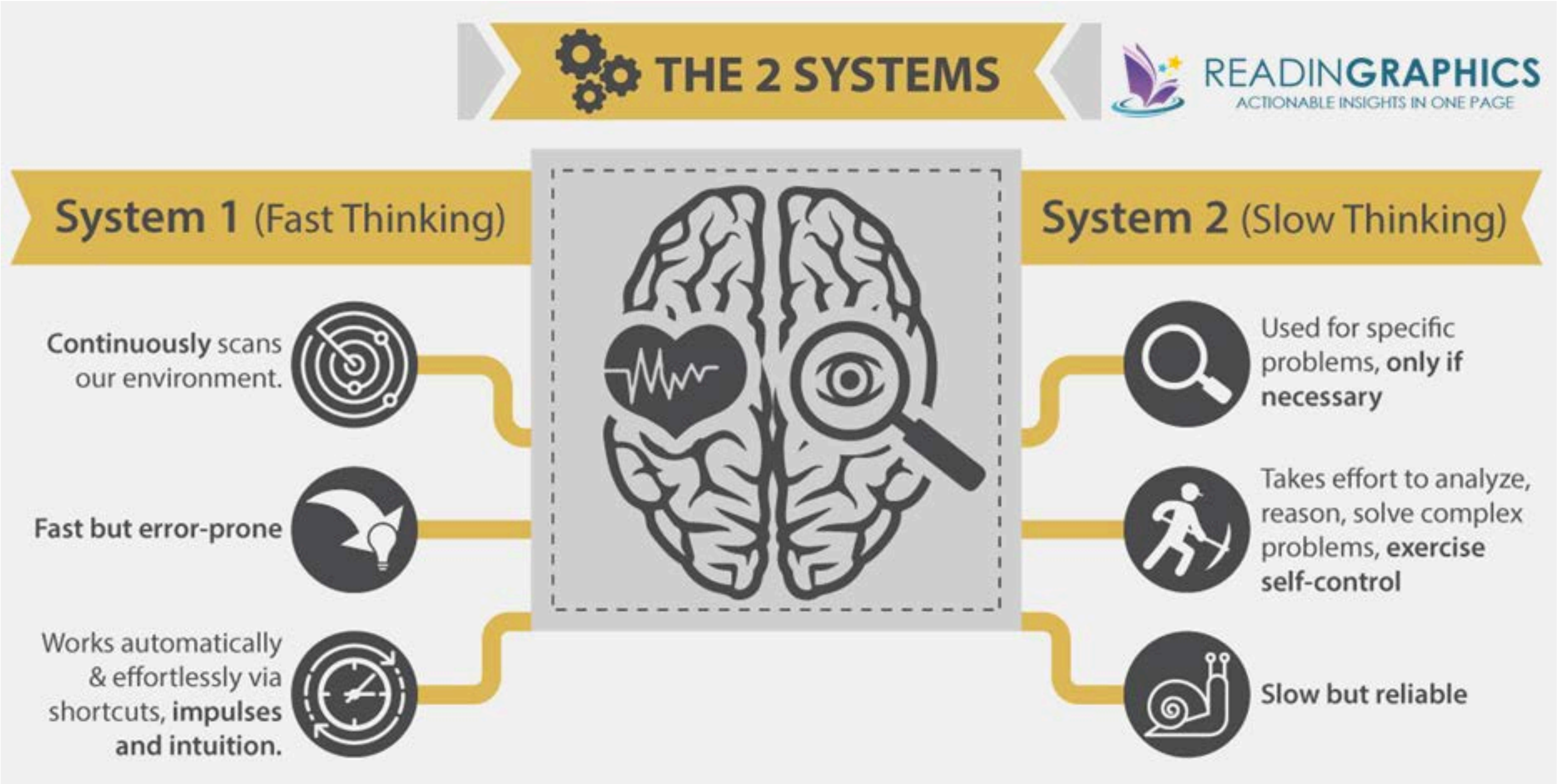
N. Kourtellis, G. De Francisci Morales, A. Bifet, A. Murdopo: "VHT: Vertical Hoeffding Tree", 2016 Big Data Conference 2016



Single attribute tracked in single node

3. Learning Fast and Slow


THINKING,
FAST AND SLOW
DANIEL
KAHNEMAN
WINNER OF THE NOBEL PRIZE IN ECONOMICS



THE 2 SYSTEMS

READINGRAPHICS
ACTIONABLE INSIGHTS IN ONE PAGE

System 1 (Fast Thinking)

Continuously scans our environment.

Fast but error-prone

Works automatically & effortlessly via shortcuts, **impulses and intuition.**

System 2 (Slow Thinking)

Used for specific problems, **only if necessary**

Takes effort to analyze, reason, solve complex problems, **exercise self-control**

Slow but reliable

Learning Fast and Slow

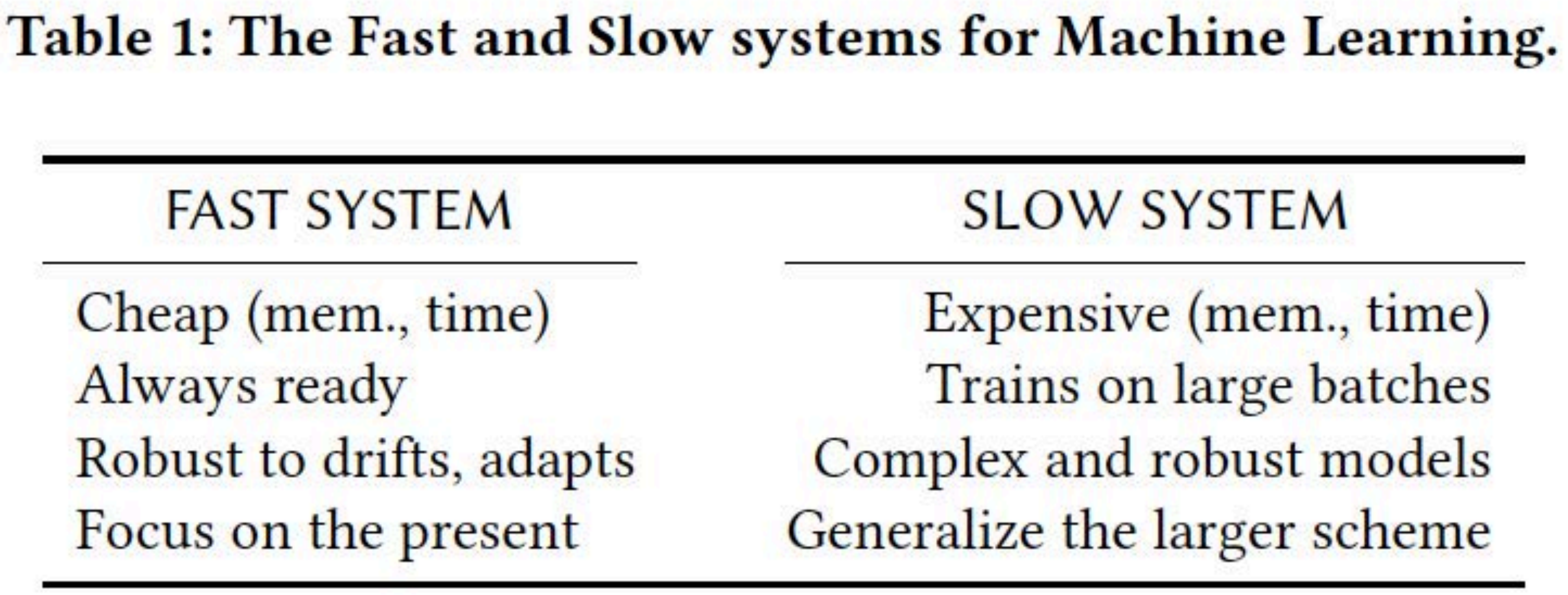
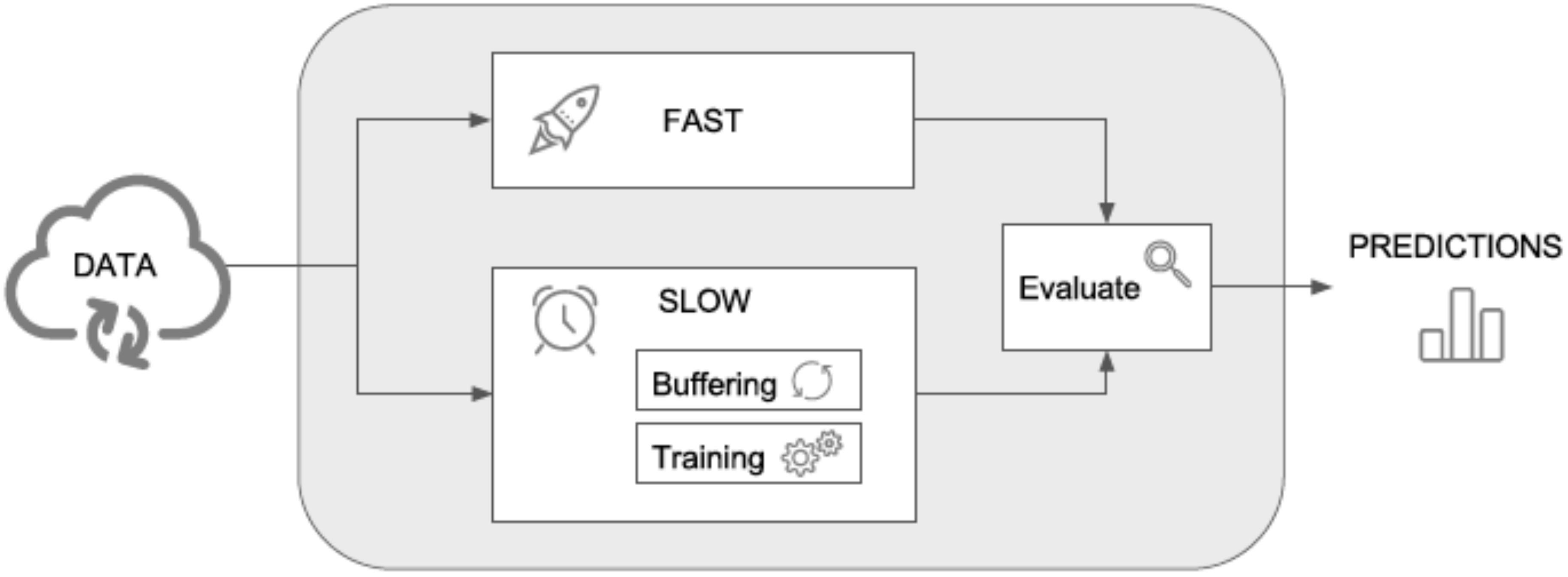


Table 1: The Fast and Slow systems for Machine Learning.

FAST SYSTEM	SLOW SYSTEM
Cheap (mem., time)	Expensive (mem., time)
Always ready	Trains on large batches
Robust to drifts, adapts	Complex and robust models
Focus on the present	Generalize the larger scheme

Learning Fast and Slow



Learning Fast and Slow

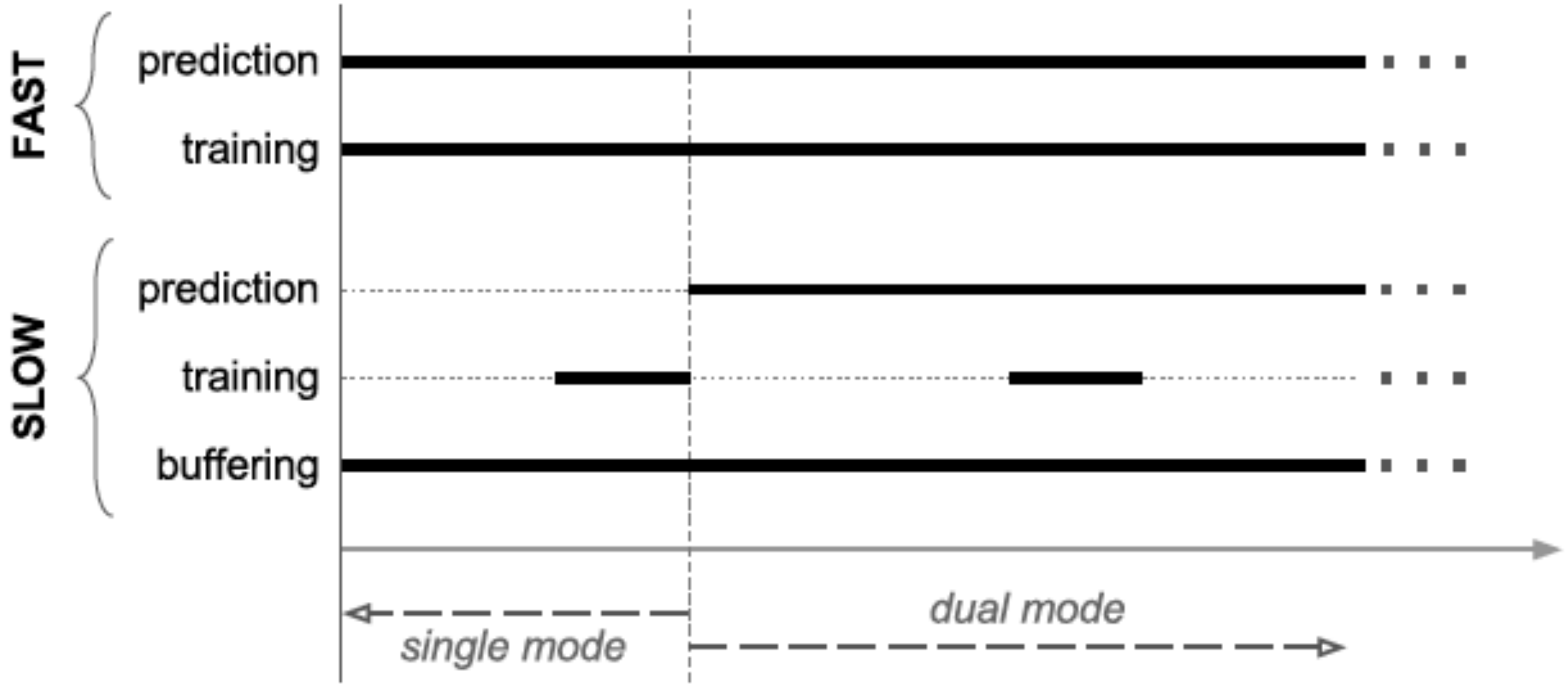


Figure 2: FSC operation modes.

Learning Fast and Slow

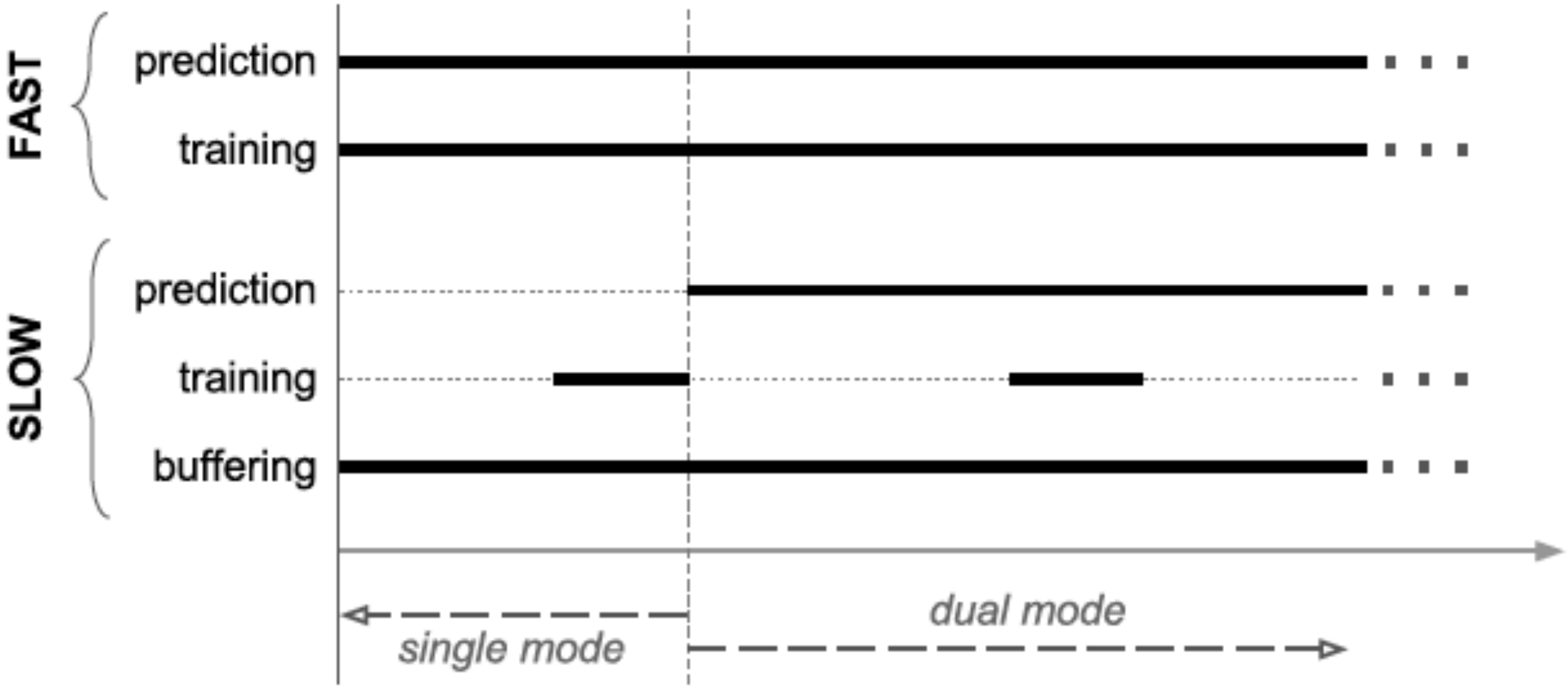
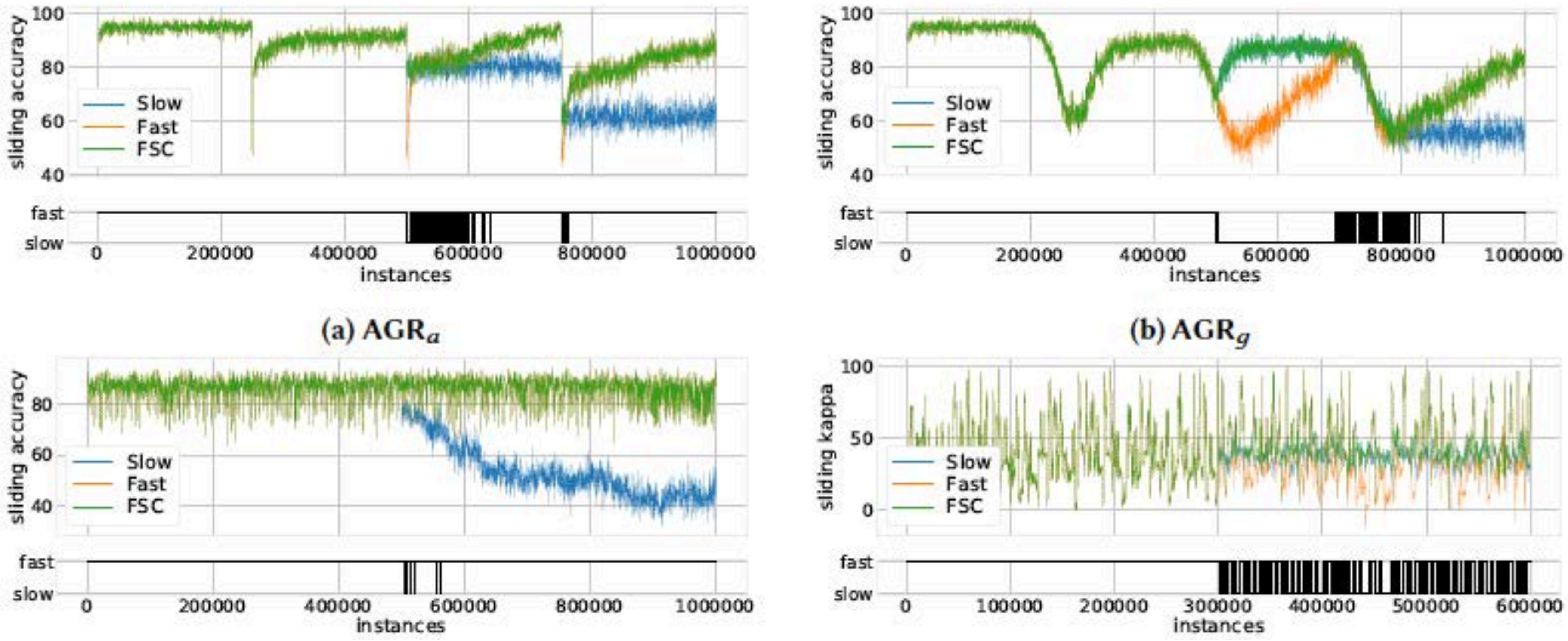


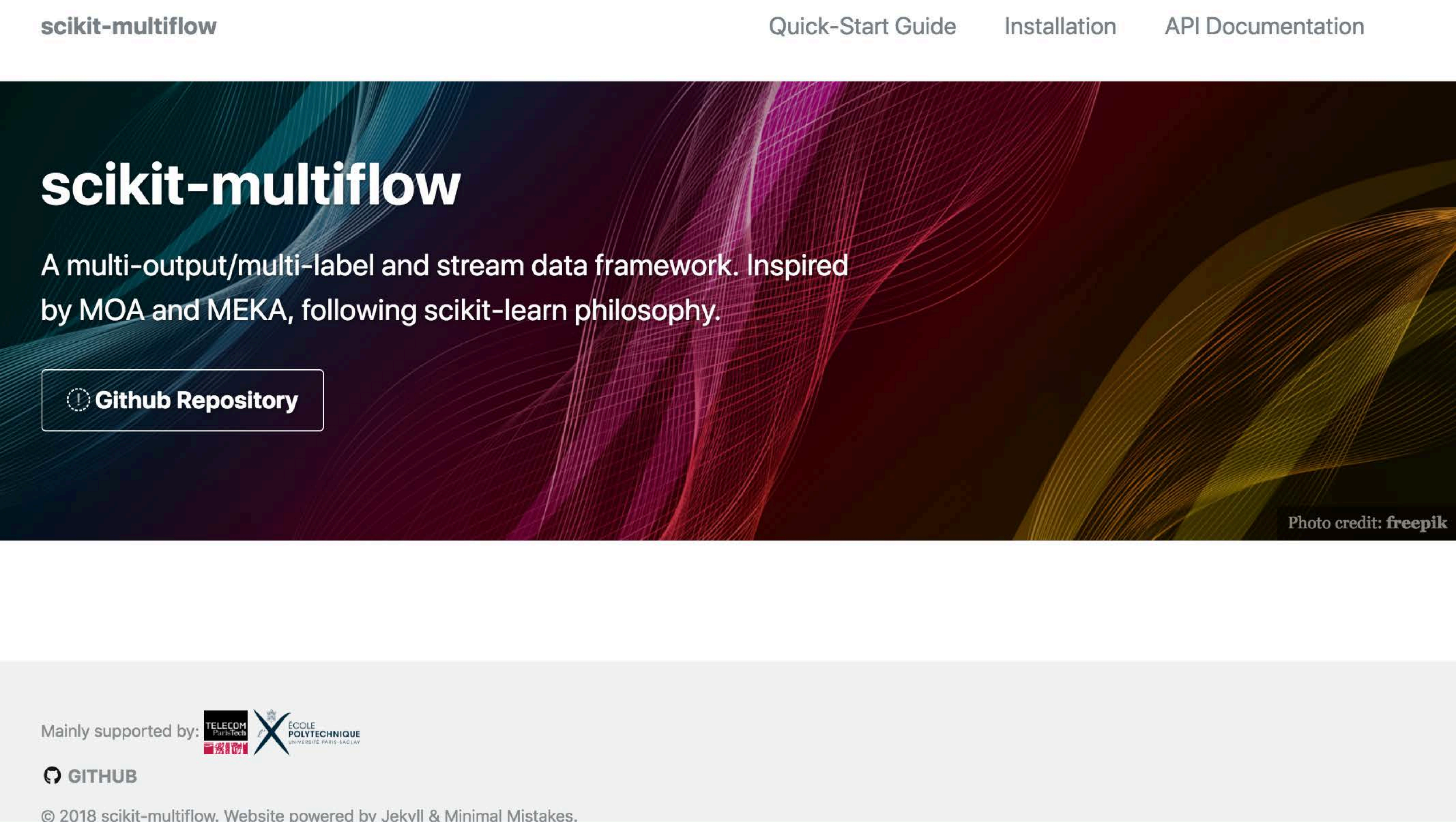
Figure 2: FSC operation modes.

Learning Fast and Slow



(a) AGR$_a$

(b) AGR$_g$



scikit-multiflow

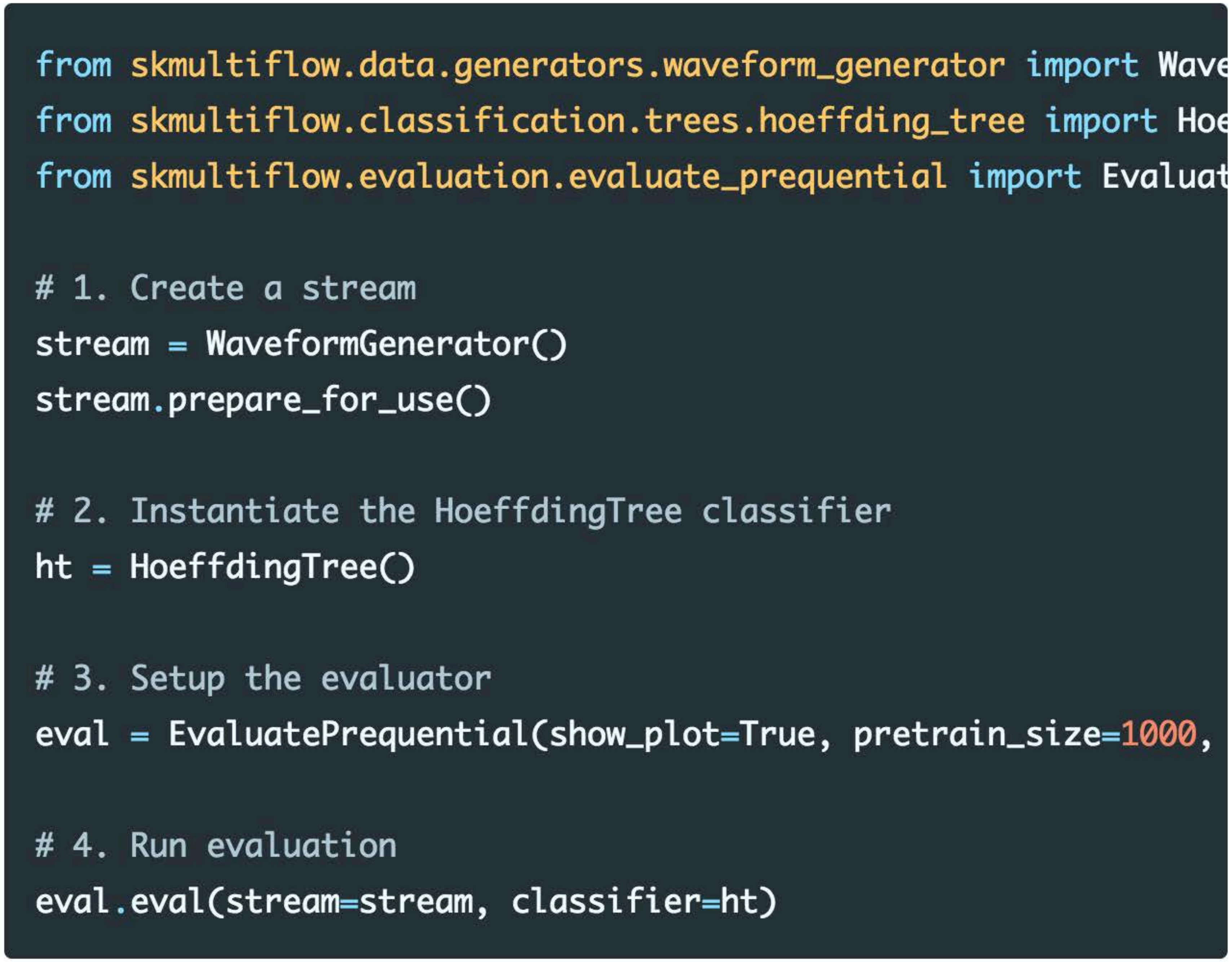
A multi-output/multi-label and stream data framework. Inspired by MOA and MEKA, following scikit-learn philosophy.

Photo credit: **freepik**

Mainly supported by:

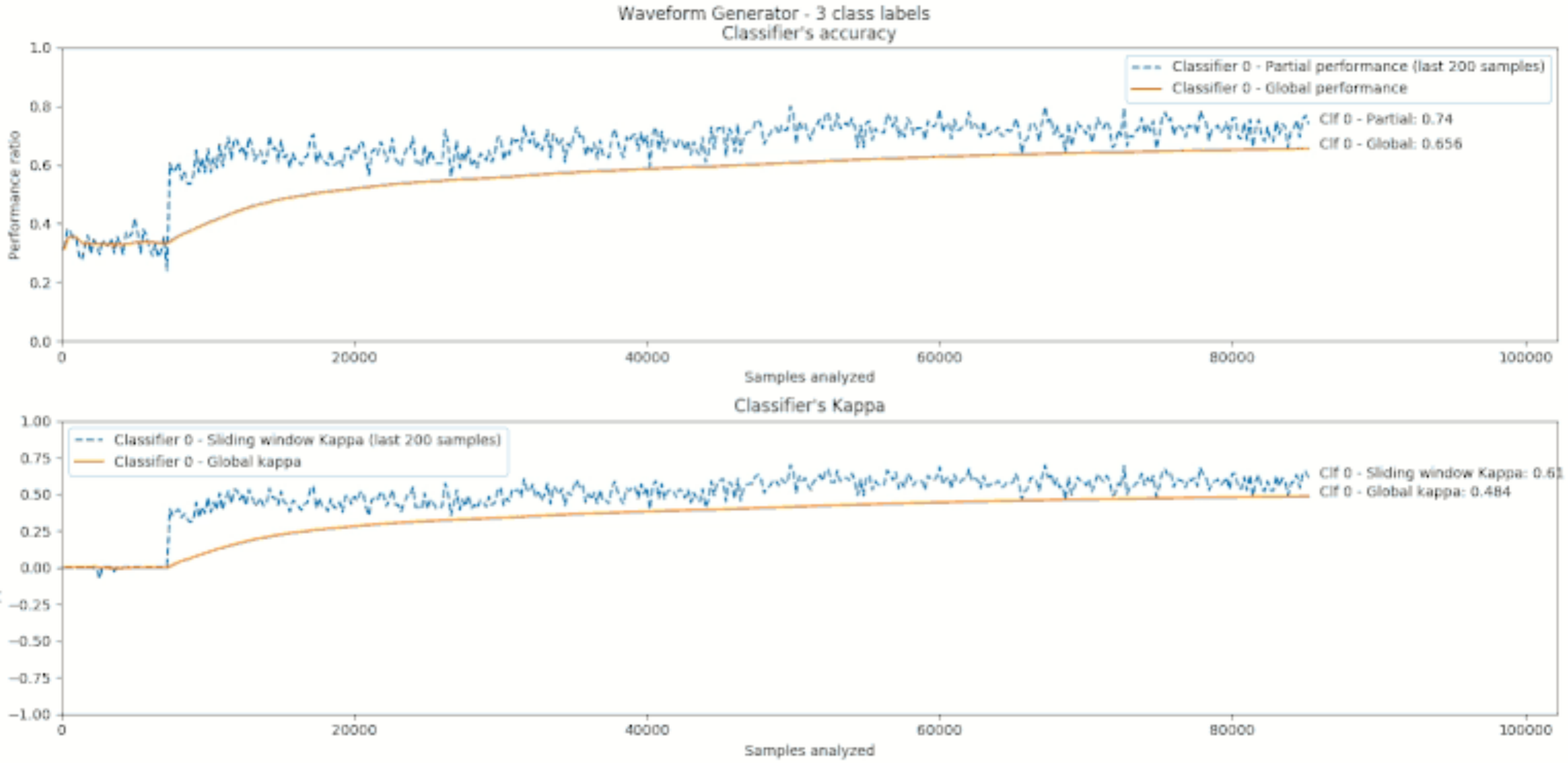
© 2018 scikit-multiflow. Website powered by Jekyll & Minimal Mistakes.

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)
```

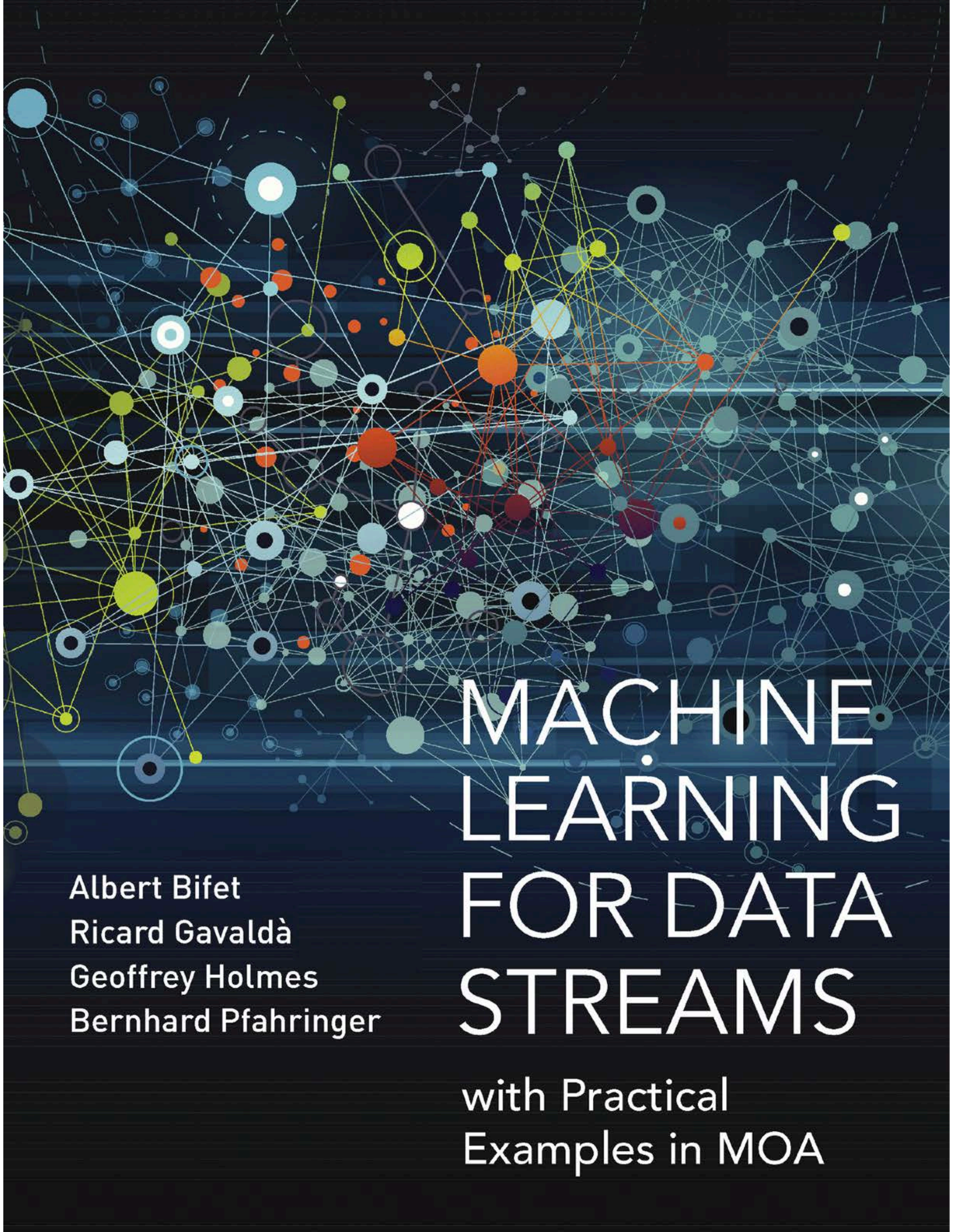
scikit-multiflow



Summary

- Green AI
- Explainable AI
- Ethical Issues
- Open AI
- Distributed Data Stream Mining
- Learning Fast and Slow

The MIT Press



MACHINE LEARNING FOR DATA STREAMS

with Practical Examples in MOA

Albert Bifet
Ricard Gavaldà
Geoffrey Holmes
Bernhard Pfahringer

Thanks!

@abifet

Machine Learning for Data Streams

Albert Bifet (@abifet)

Cisco-Ecole Polytechnique Symposium 2018,
10 April 2018

TELECOM ParisTech

l'X ÉCOLE POLYTECHNIQUE UNIVERSITÉ PARIS-SACLAY