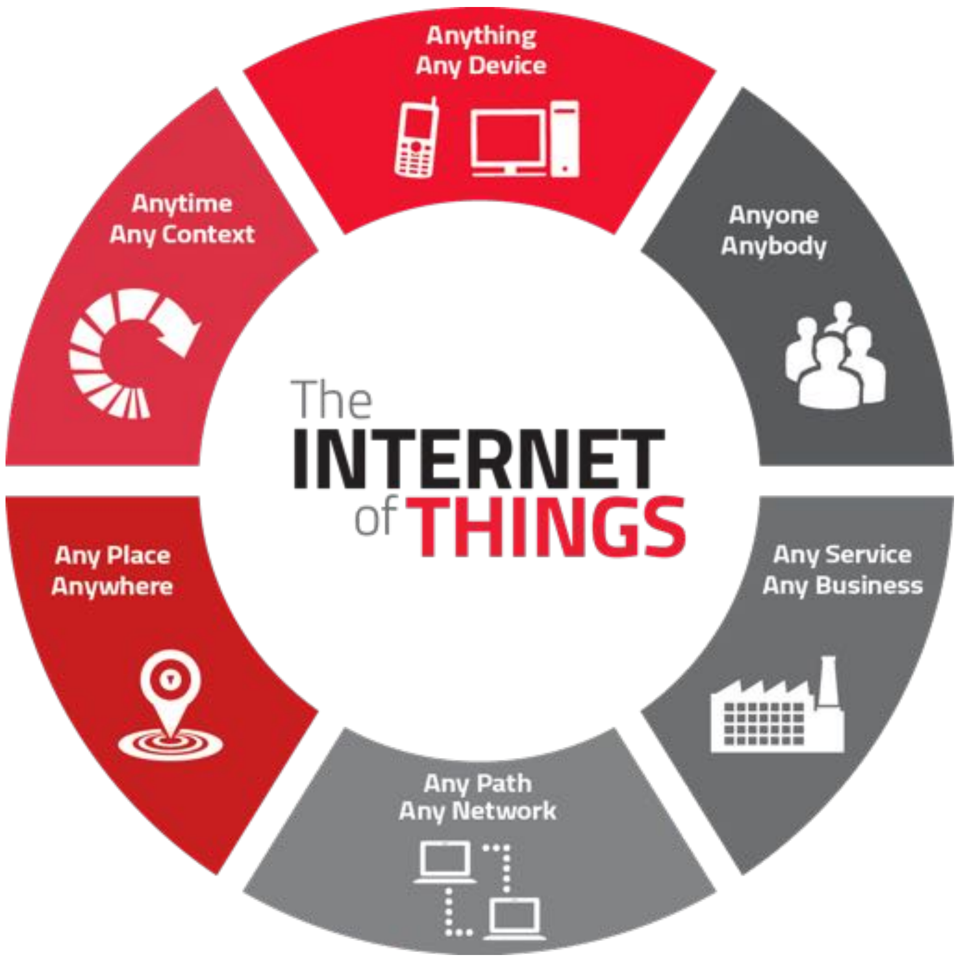
Machine Learning for Data Streams

Albert Bifet (@abifet)

Paris
TMA Conference 2019
20 June 2019

TELECOM Paris
IP PARIS
INSTITUT POLYTECHNIQUE DE PARIS
THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

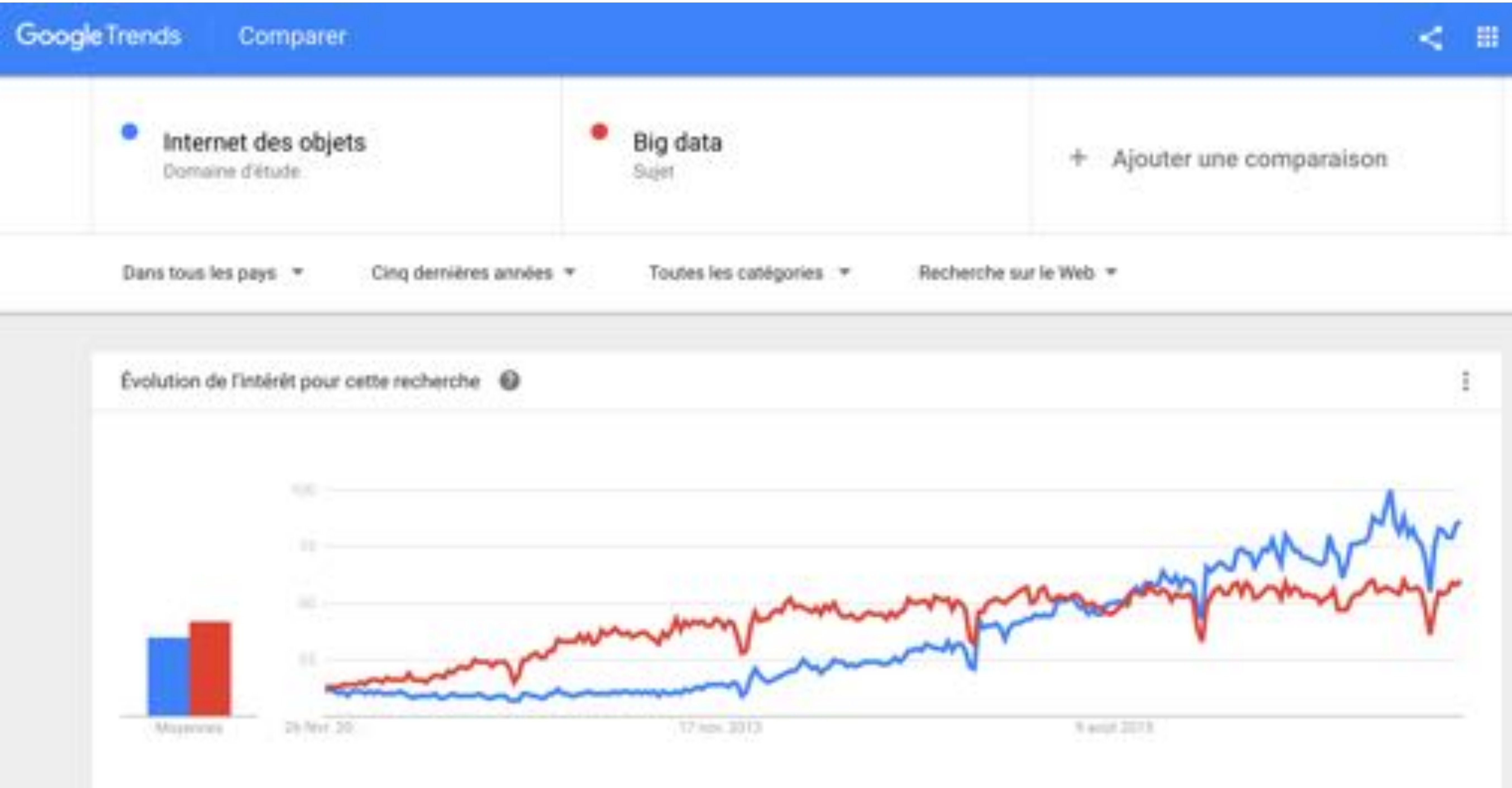
Internet of Things



IoT: sensors and actuators connected by networks to computing systems.

- Gartner predicts 20.8 billion IoT devices by 2020.
- IDC projects 32 billion IoT devices by 2020

IoT versus Big Data



AI/Machine Learning is the new Electricity

- **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.

“Over the past decades, computers have broadly automated tasks that programmers could **describe with clear rules and algorithms.**

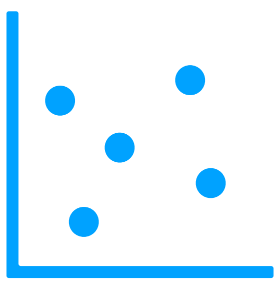
Modern machine learning techniques now allow us to do the same for tasks where **describing the precise rules is much harder.**”

–Jeff Bezos

“Machine Learning is a way of solving problems without knowing how to solve them”

–Unknown Author

Computer Science



Data

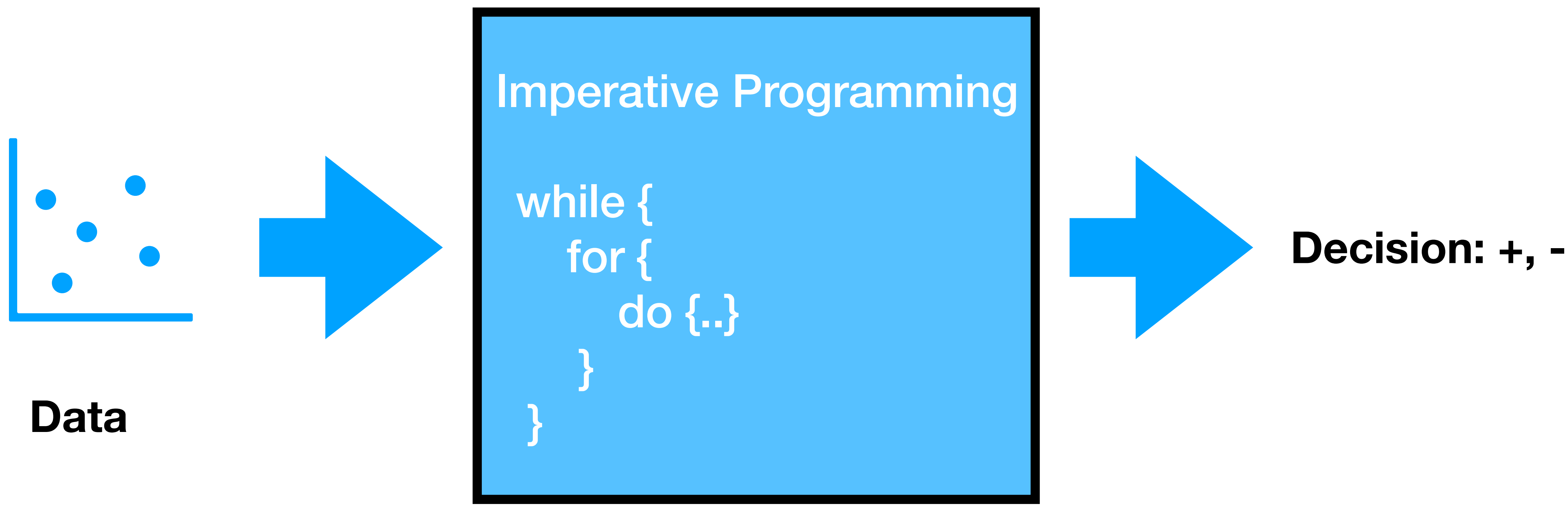
Imperative Programming

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

Decision: +, -

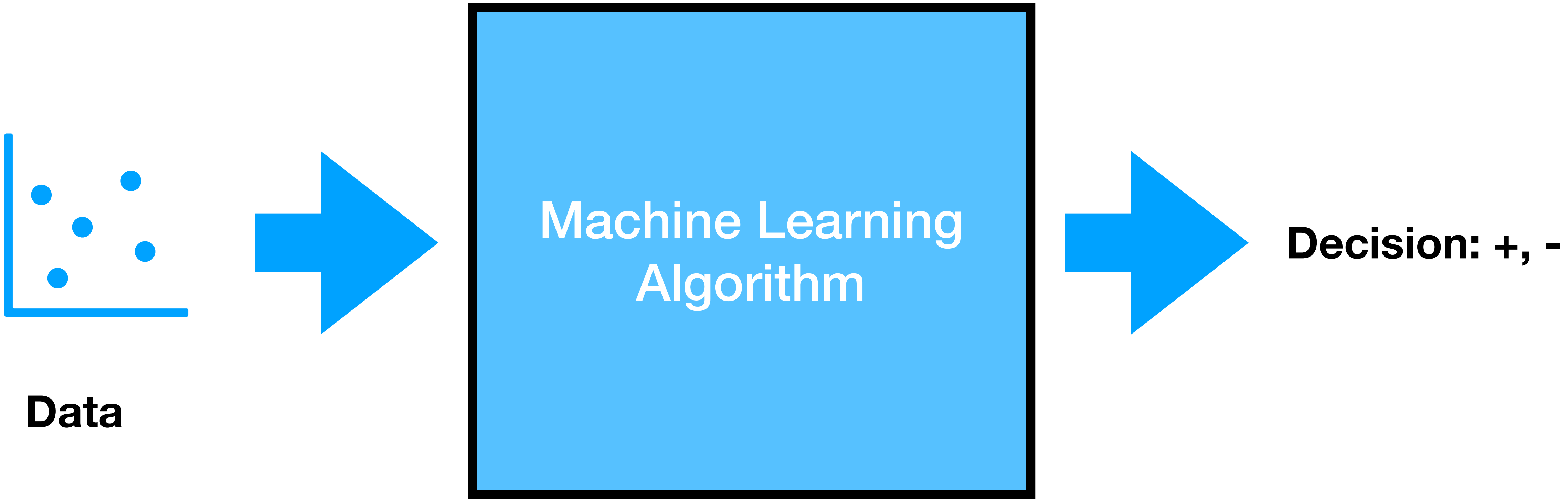
Software 1.0

Computer Science



Software 1.0

Machine Learning

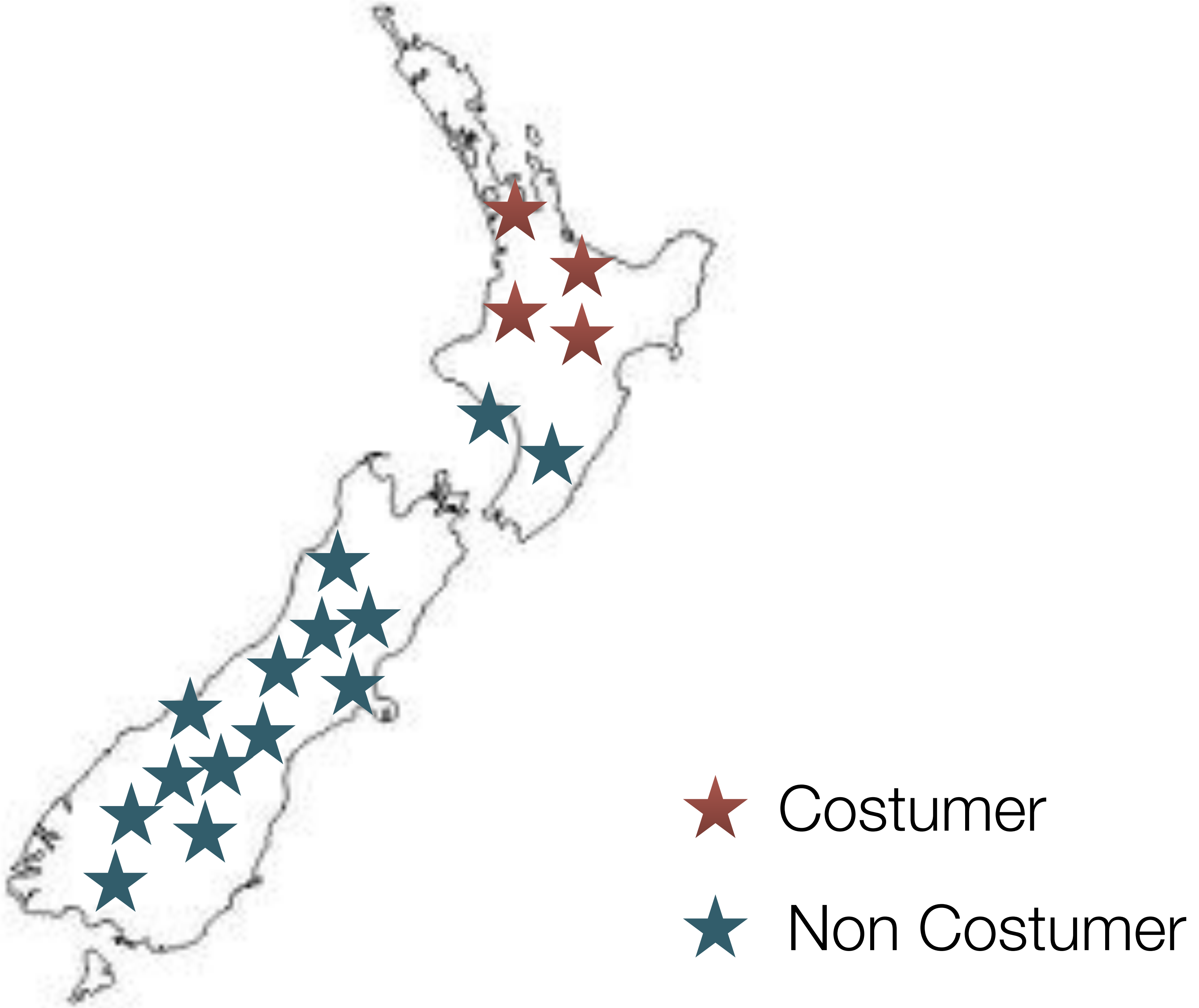


Software 2.0

5 Minutes Course on Machine Learning

AI/ML = Data + **Algorithms** + Computing Power

Classification of Visitors to a Website


Costumer
Non Costumer


Is this visitor
a costumer?
Costumer
Non Costumer

Classification

Majority Class

★ Non Costumer

Is this visitor a costumer?

Classification

Linear Classifier

★ Costumer

Classification

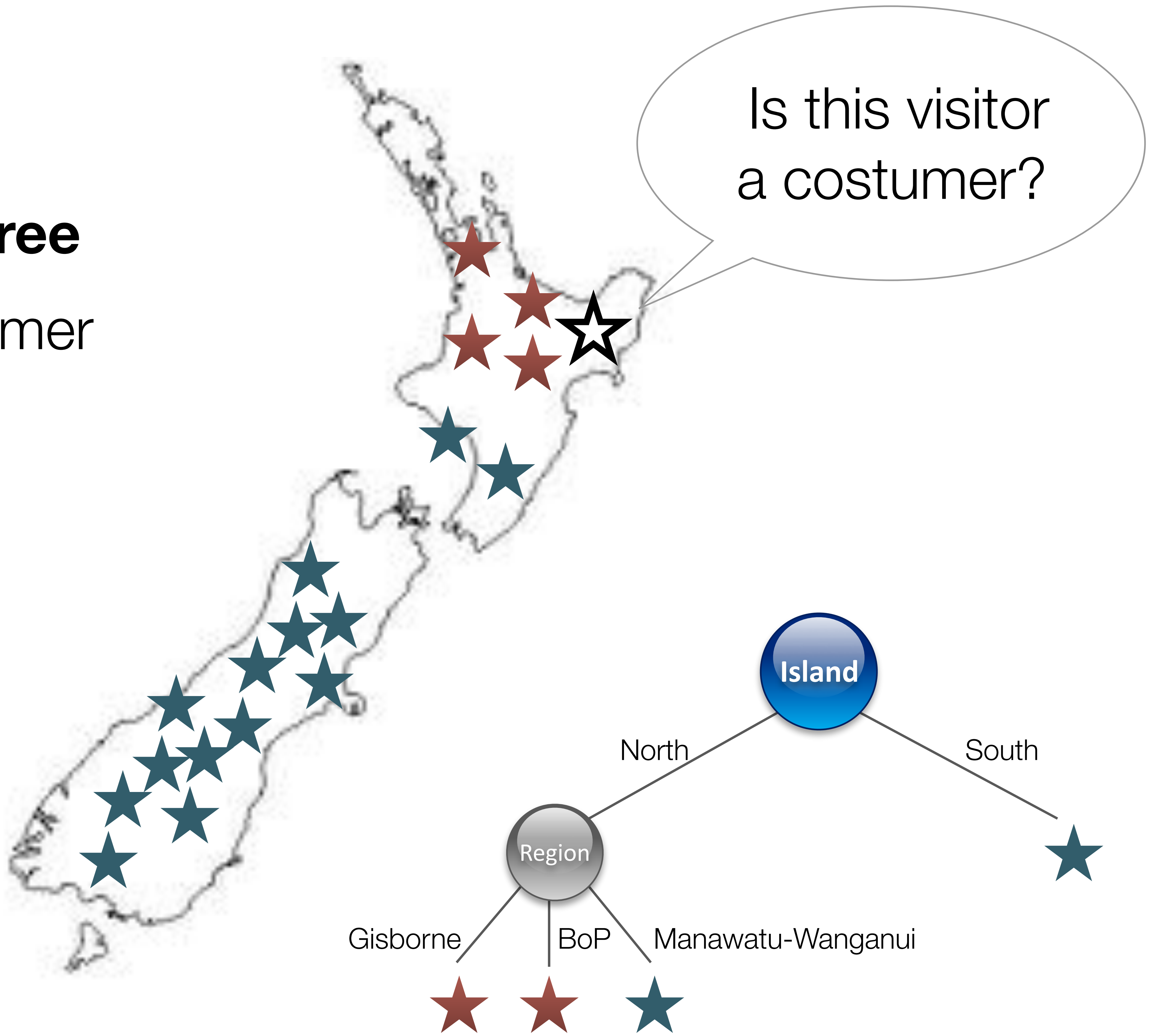
Nearest Neighbour

★ Costumer

Classification

Decision Tree

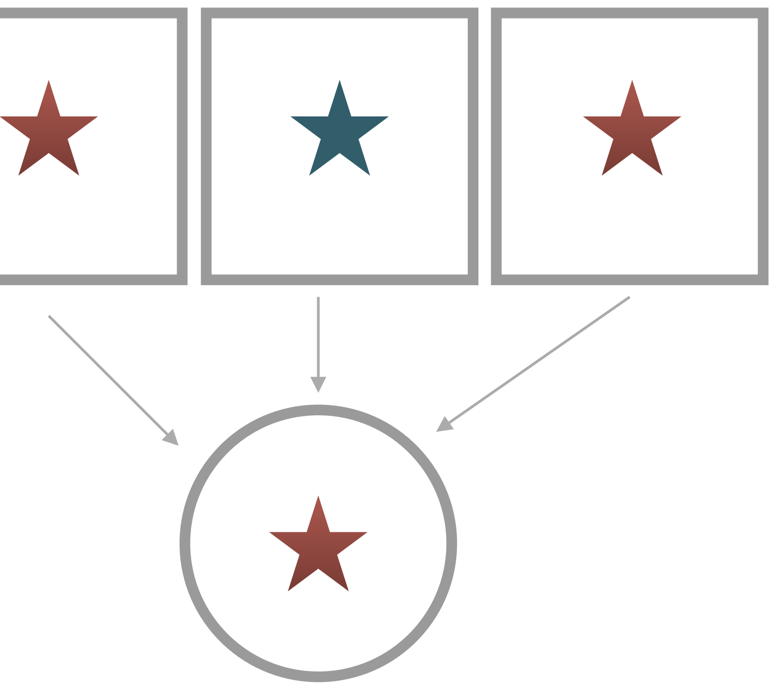
★ Costumer



Classification

Random Forest: Ensemble of Random Trees

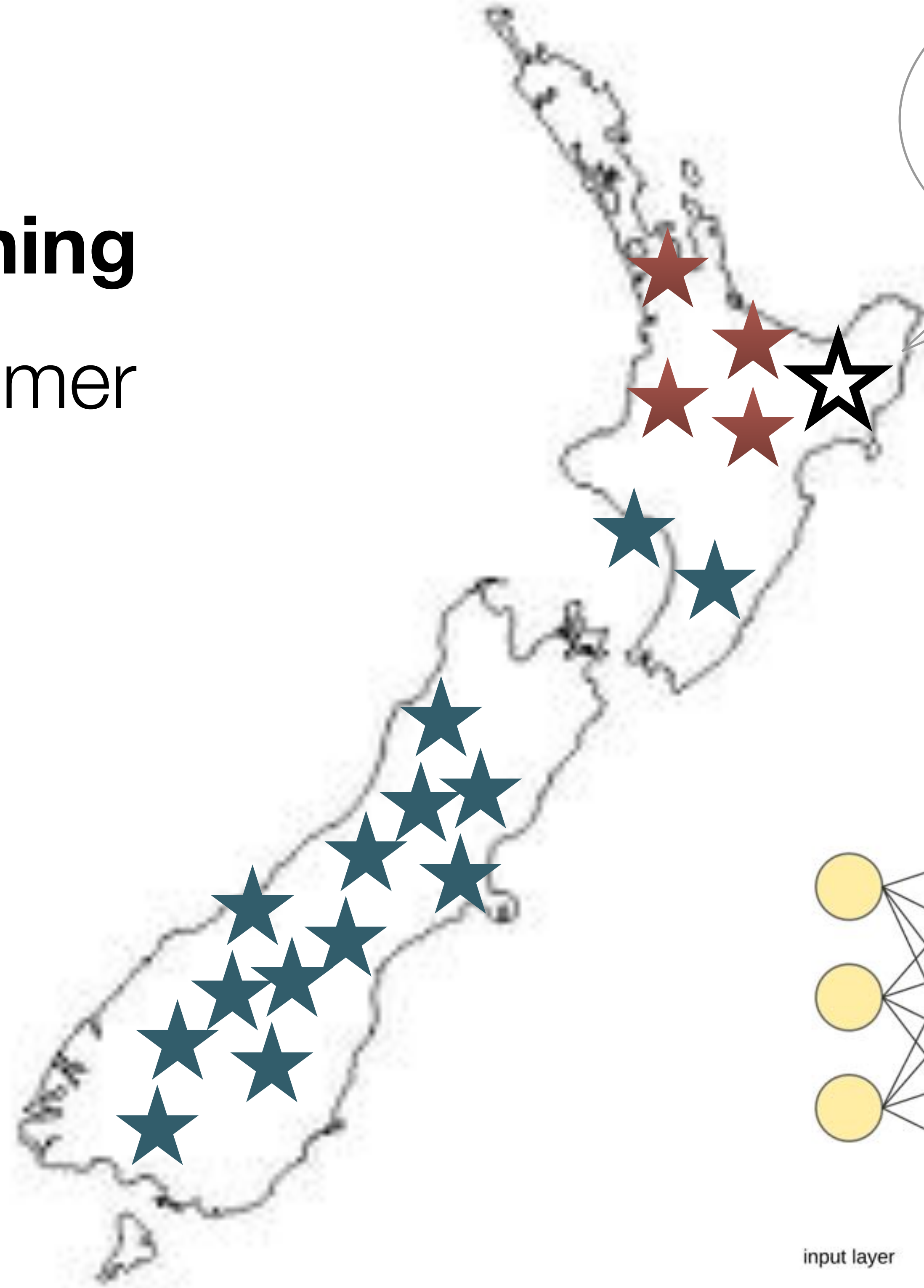
★ Costumer



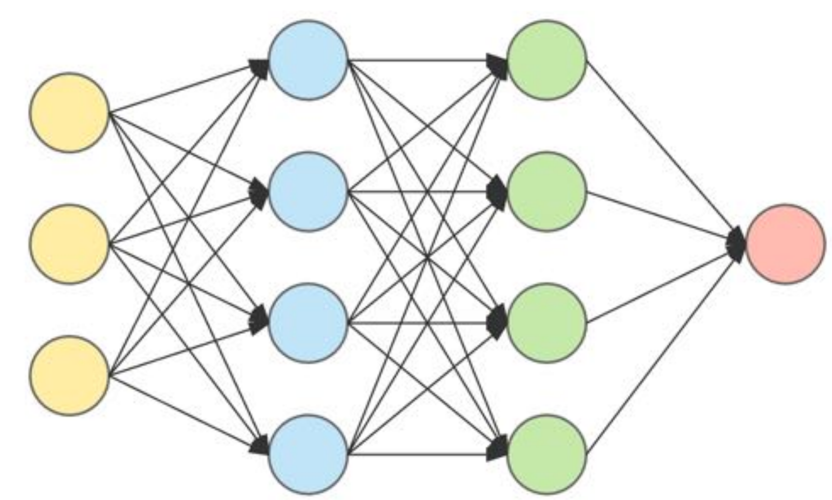
Classification

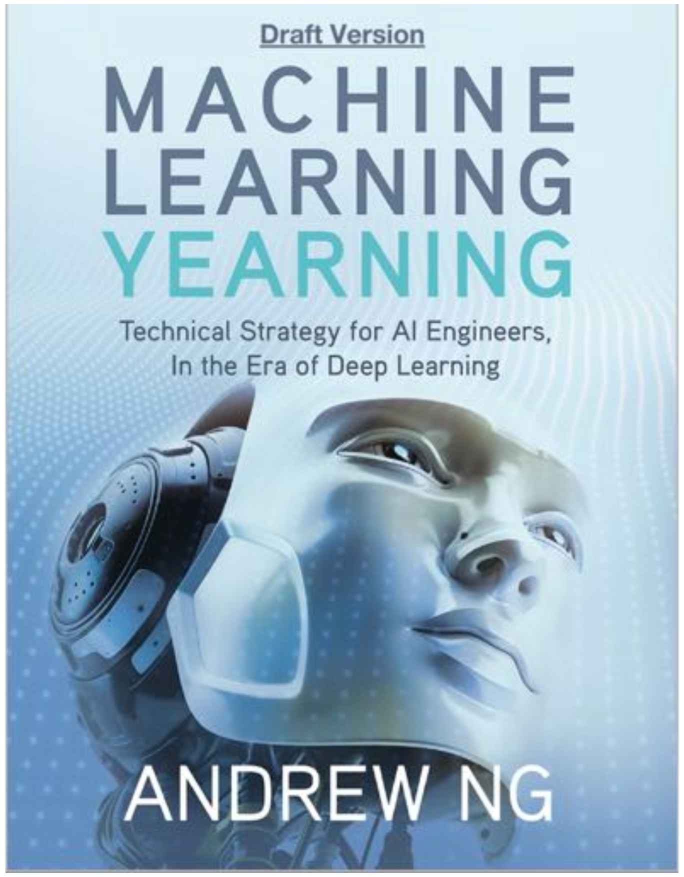
Deep Learning

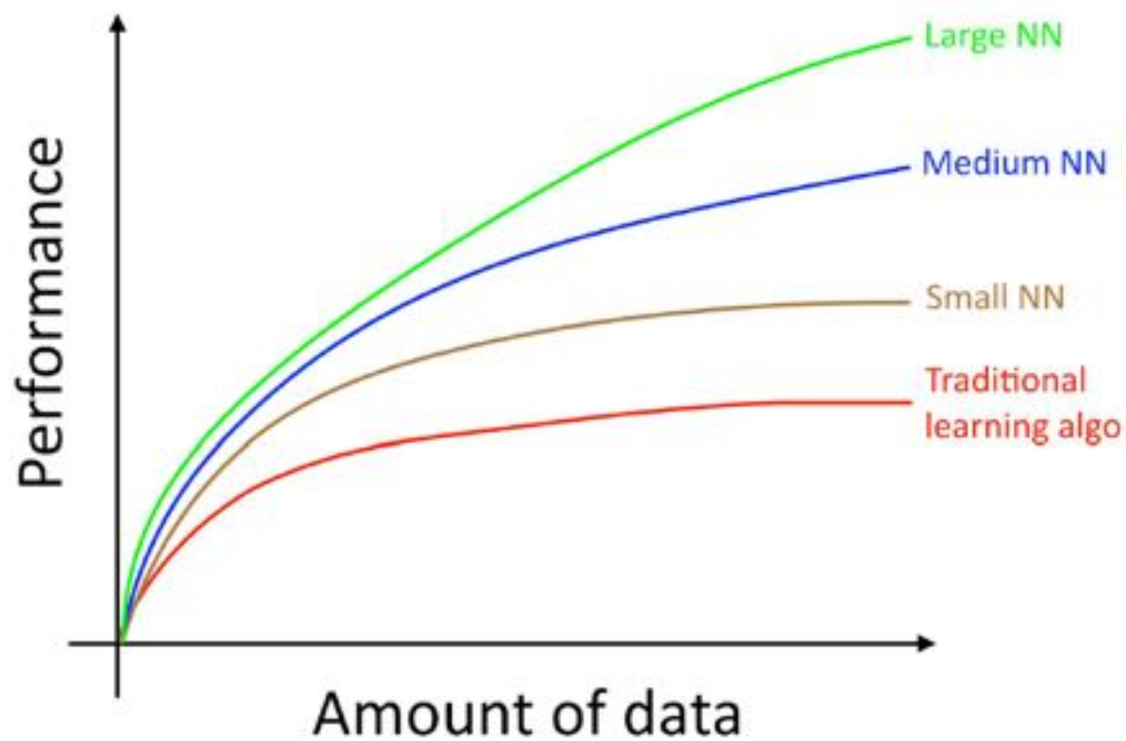
★ Costumer



Is this visitor
a costumer?







Scale drives machine learning progress

Many of the ideas of deep learning (neural networks) have been around for decades. Why are these ideas taking off now?

Two of the biggest drivers of recent progress have been:

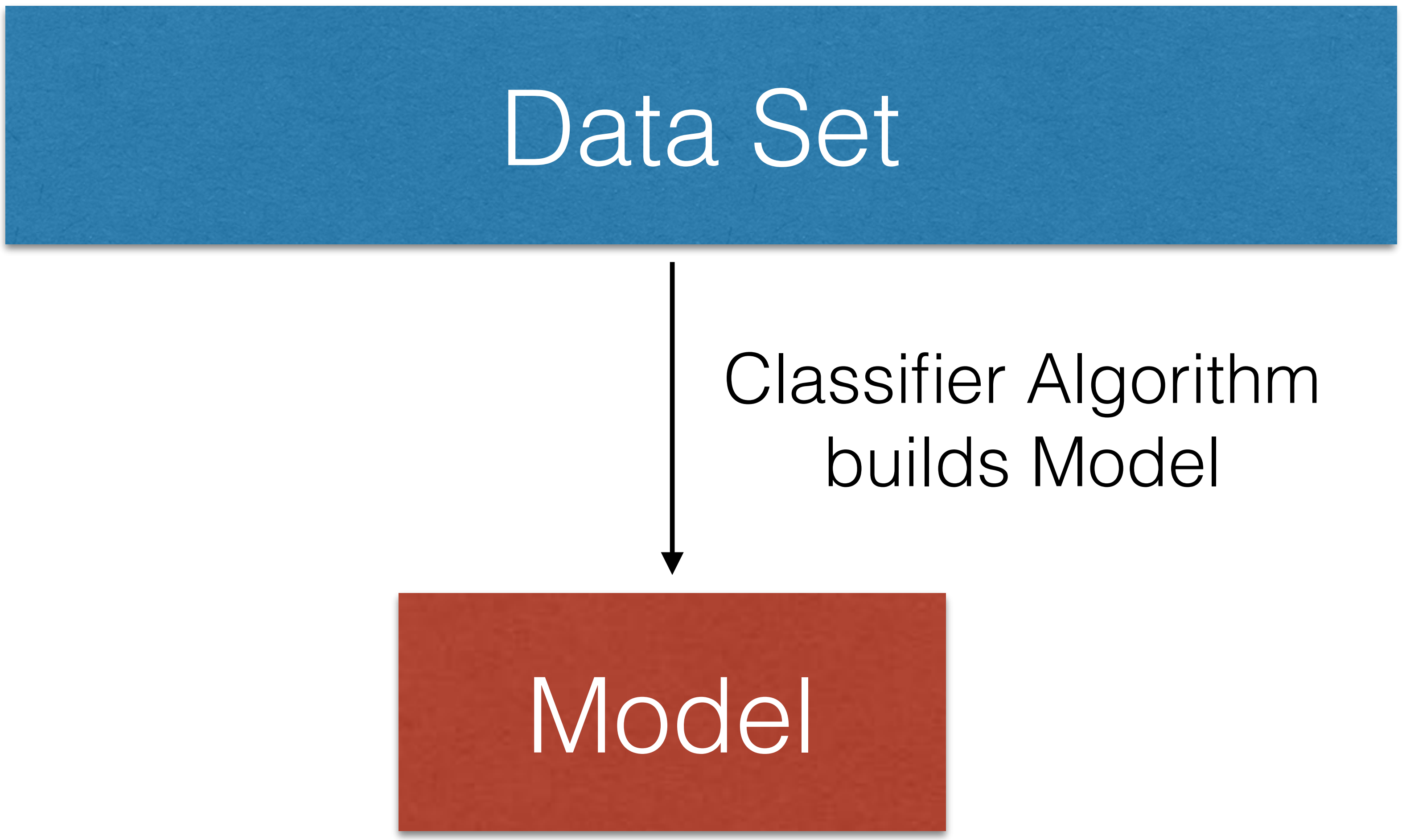
- **Data availability.** People are now spending more time on digital devices (laptops, mobile devices). Their digital activities generate huge amounts of data that we can feed to our learning algorithms.
- **Computational scale.** We started just a few years ago to be able to train neural networks that are big enough to take advantage of the huge datasets we now have.

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.

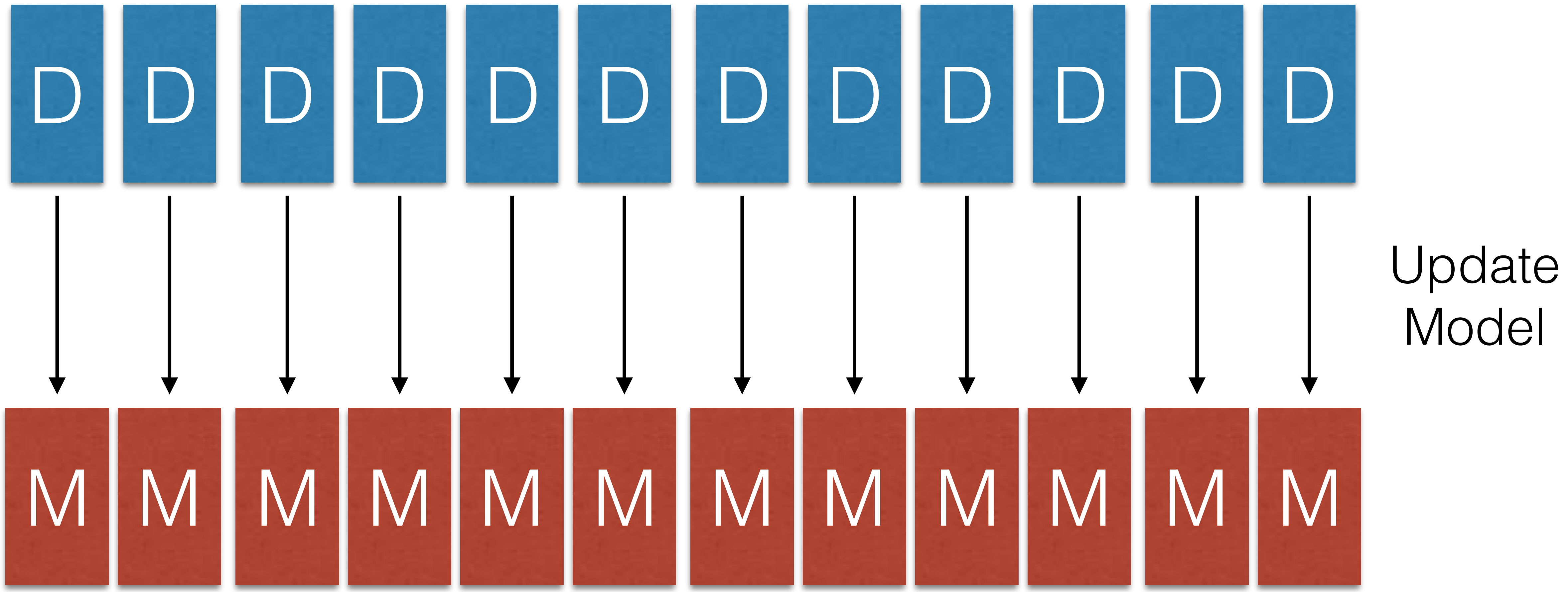
Data Streams

- 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

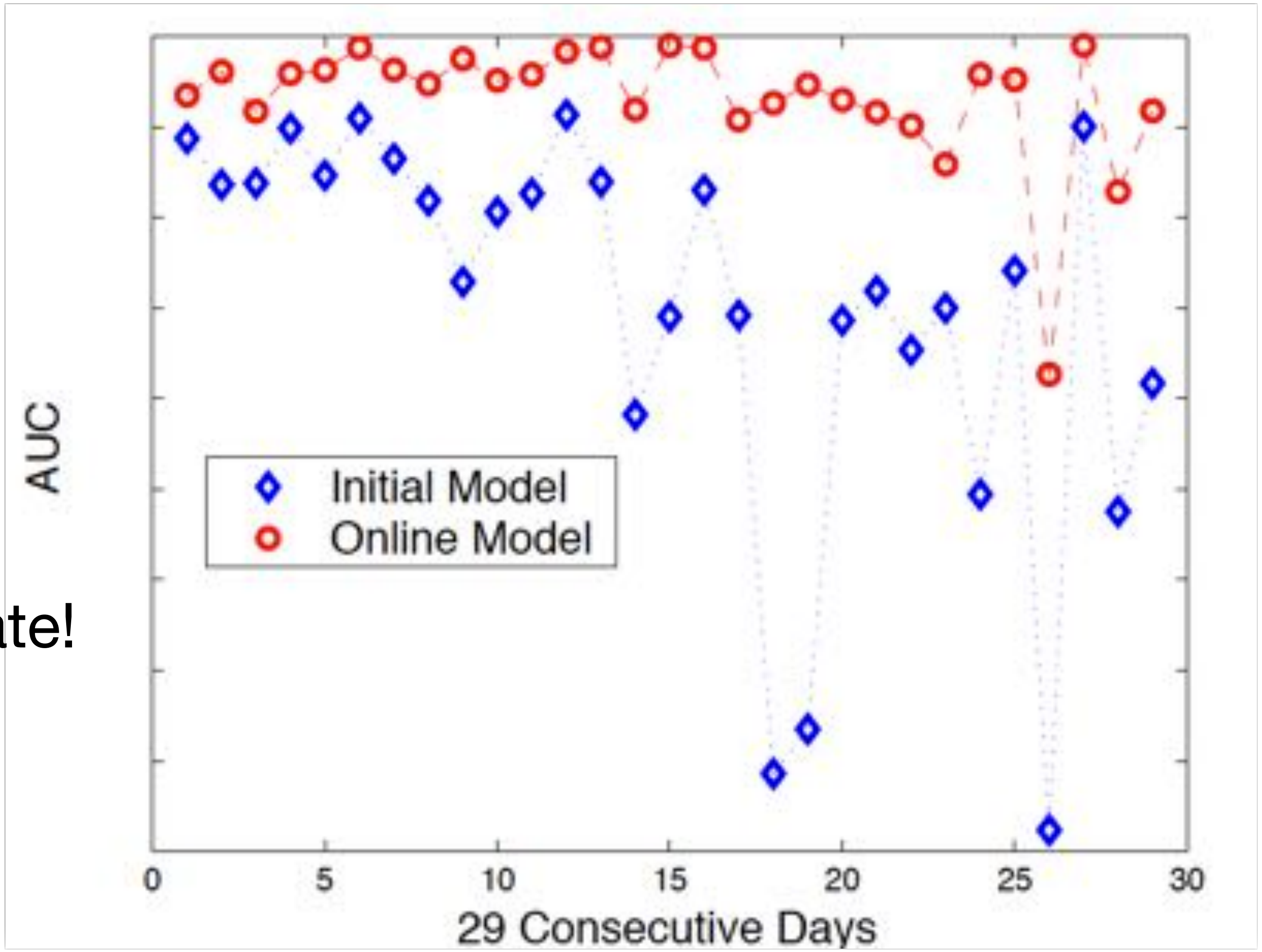


Data Stream Approach

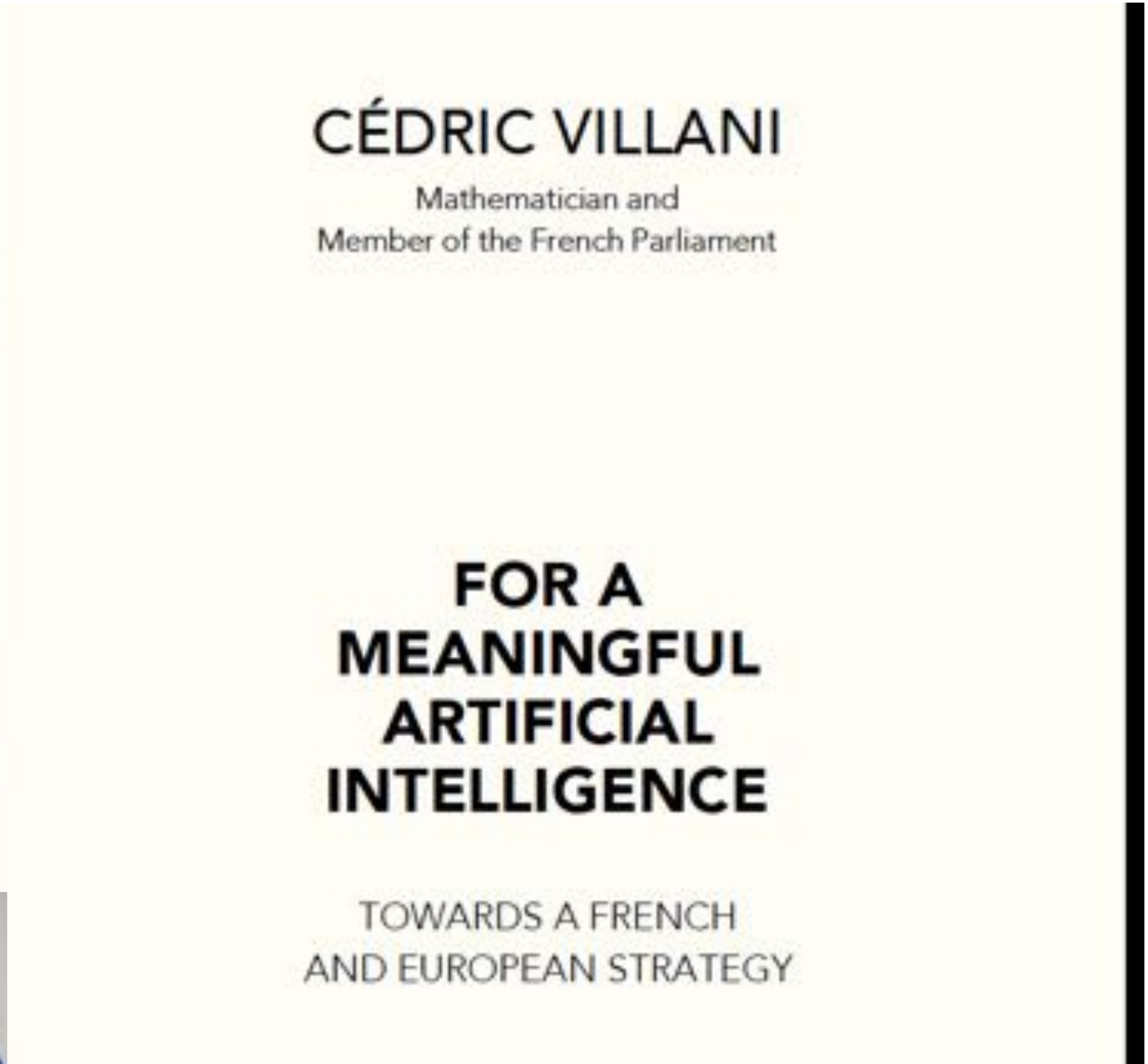
Infinite training sets
Dynamic models

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

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

MOA

Richard Kirkby

Software Developer at
11Ants Analytics Ltd

Geoff Holmes

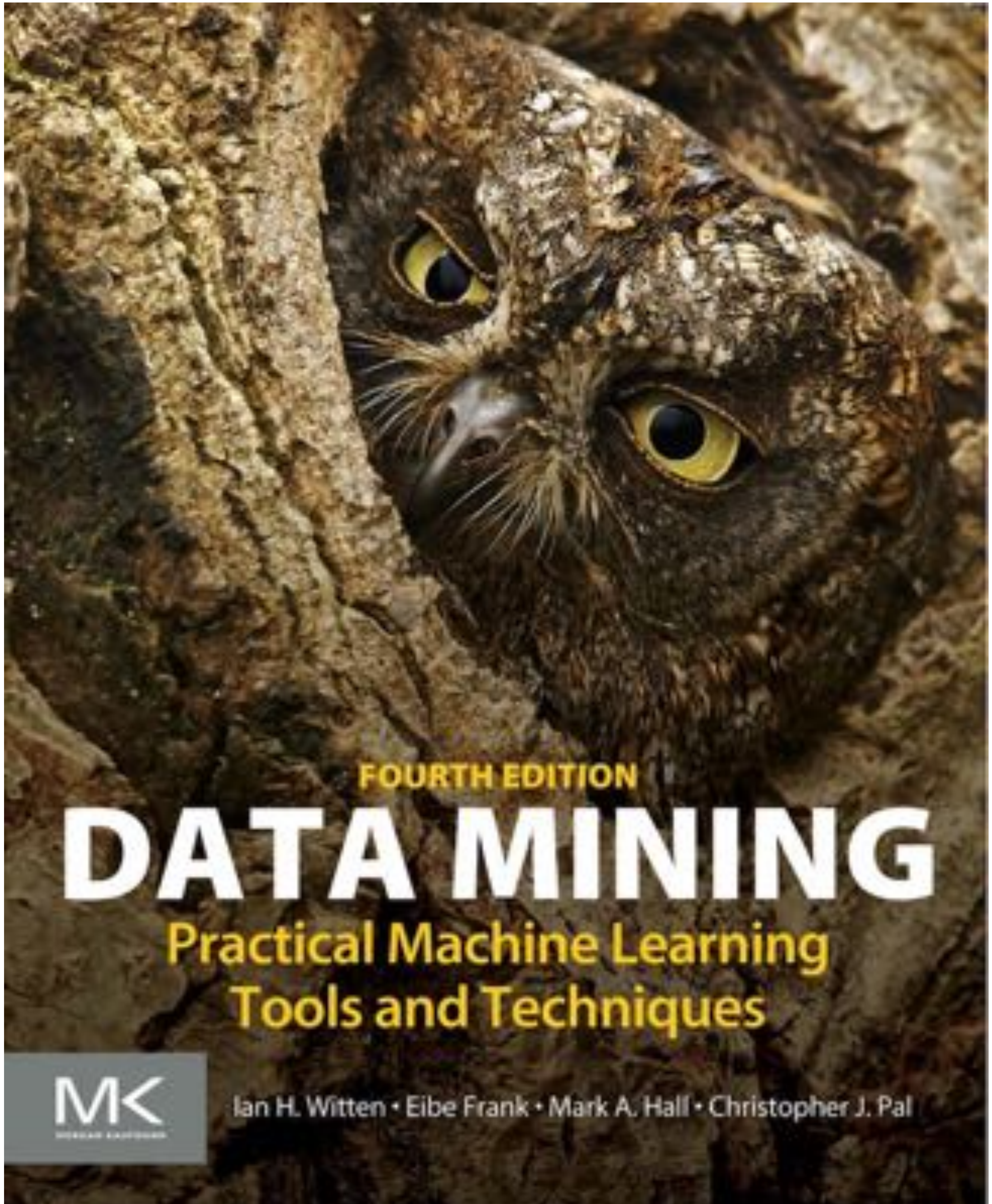
Dean of Computing & Mathematical Sciences
University of Waikato

Bernhard Pfahringer

Computer Science Department
University of Waikato

Main Contributors

- **Weka ML Group**: Peter Reutemann, Eibe Frank, Mike Mayo
- Jesse Read, Indrė Žliobaitė, Philipp Kranen, Hardy Kremer, Timm Jansen, Marwan Hassani, Thomas Seidl, Dimitris Georgiadis, Anastasios Gounaris, Apostolos N. Papadopoulos, Kostas Tsichlas, Yannis Manolopoulos, Dariusz Brzeziński, Ricard Gavaldà, Alex Catarineu, Joao Gama, Ricardo Sousa, Joao Duarte, Aljaž Osojnik, …


FOURTH EDITION
DATA MINING
Practical Machine Learning
Tools and Techniques
Ian H. Witten • Eibe Frank • Mark A. Hall • Christopher J. Pal

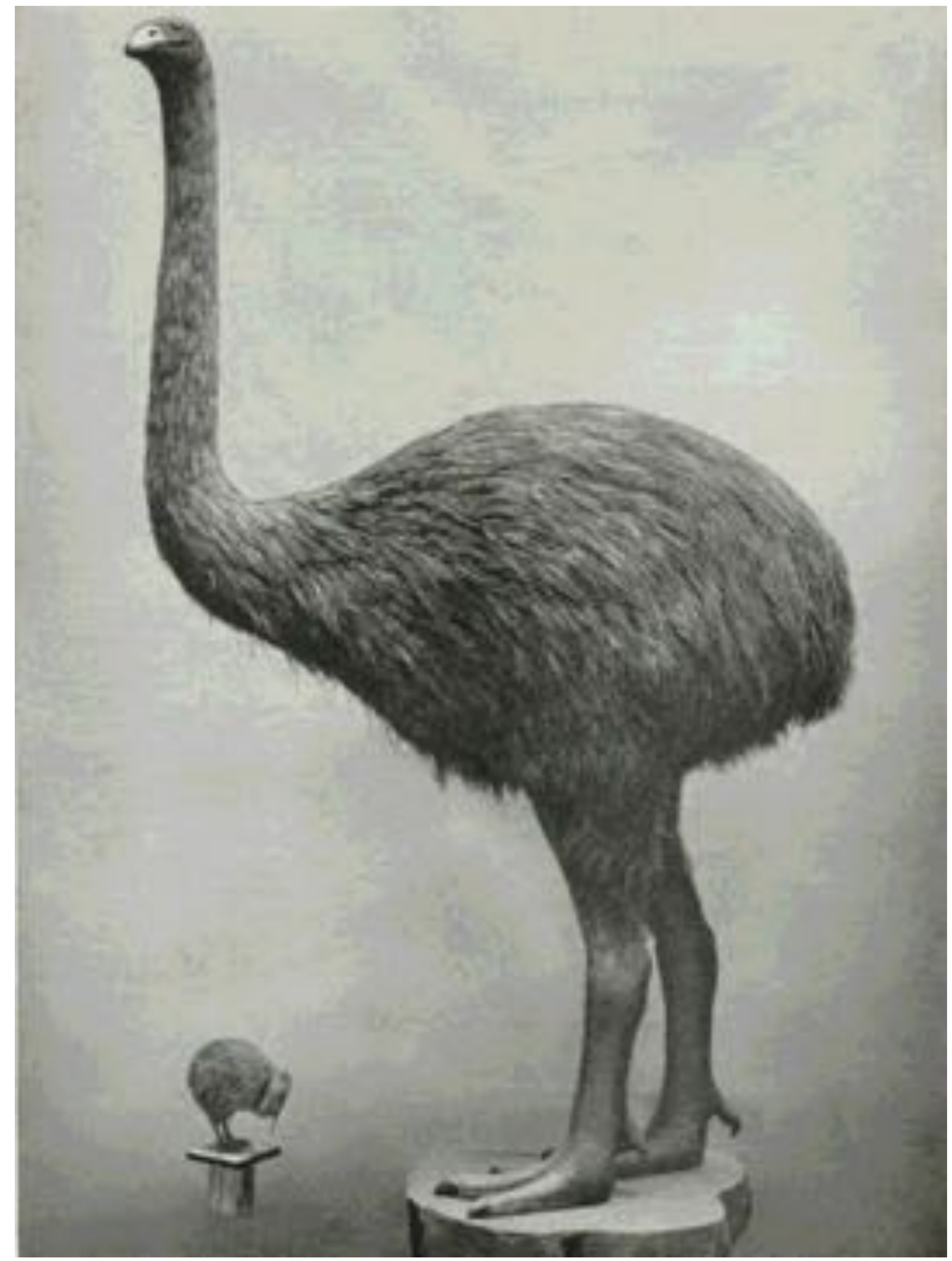
WEKA: the bird

MOA: the bird

The Moa (another native NZ bird) is not only flightless, like the Weka, but also extinct.

MOA: the bird

The Moa (another native NZ bird) is not only flightless, like the Weka, but also extinct.

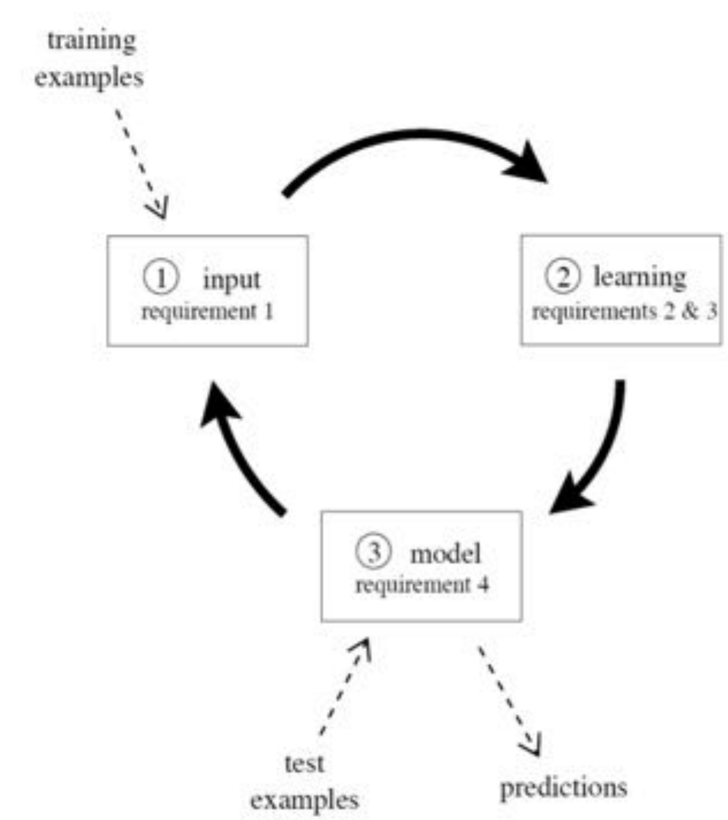


MOA: the bird

The Moa (another native NZ bird) is not only flightless, like the Weka, but also extinct.

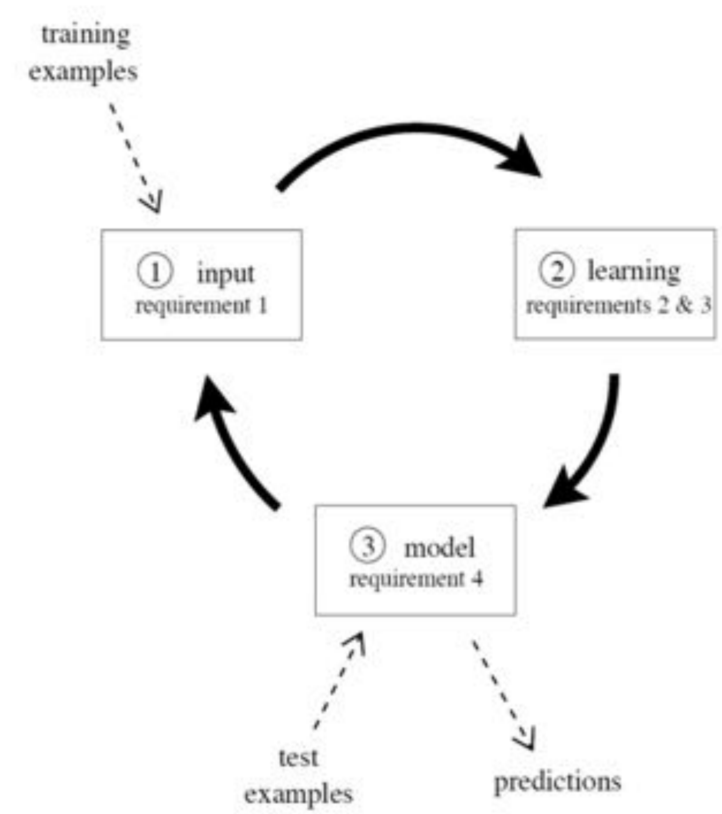
Stream Setting

- Process an example at a time,and inspect it only once (at most)
- Use a limited amount of memory
- Work in a limited amount of time
- Be ready to predict at any point



Stream Evaluation

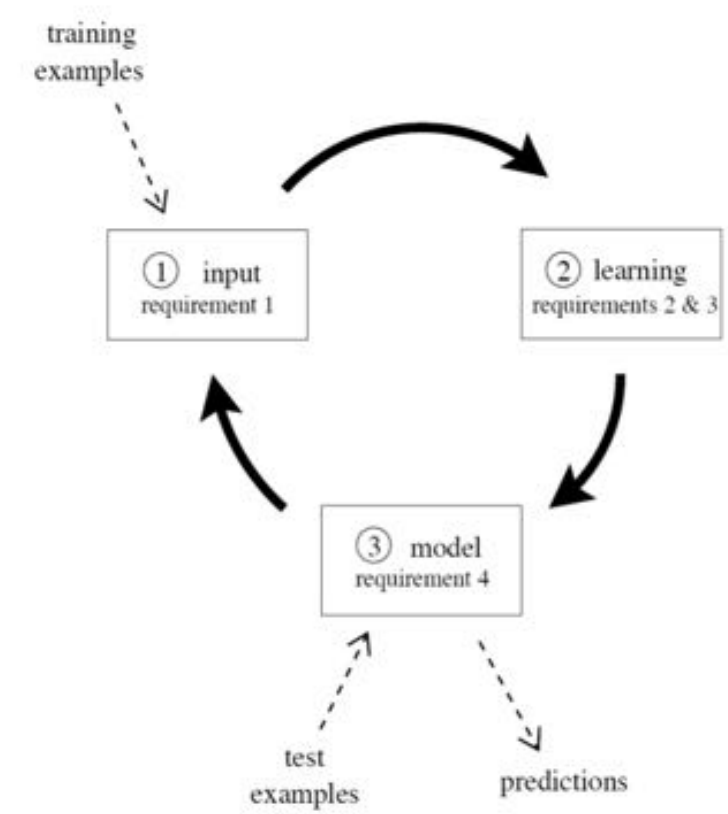
- Holdout Evaluation
- Interleaved Test-Then-Train or Prequential



Stream Evaluation

Holdout an independent test set

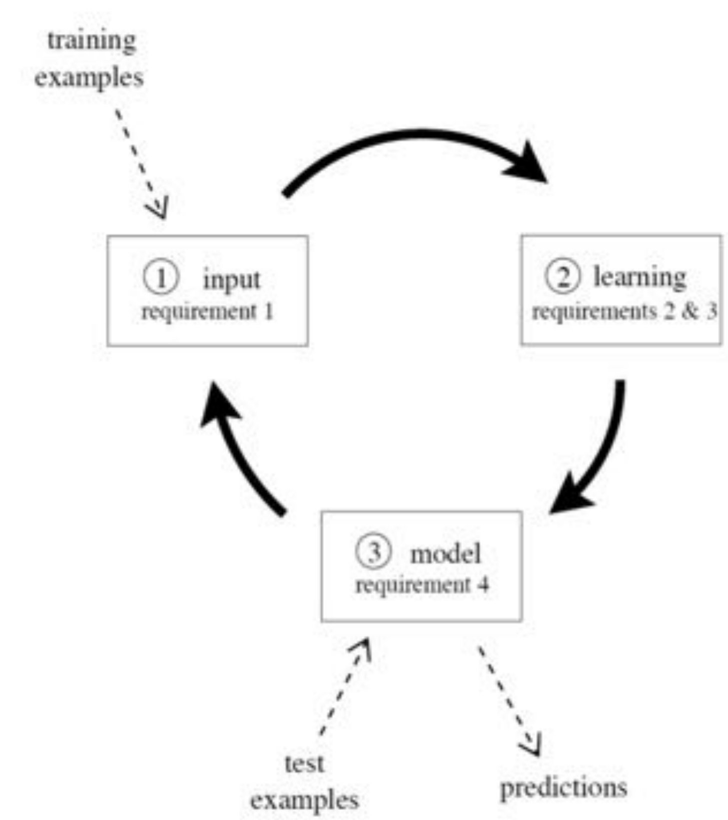
- Apply the current decision model to the test set, at regular time intervals
- The loss estimated in the holdout is an unbiased estimator



Stream Evaluation

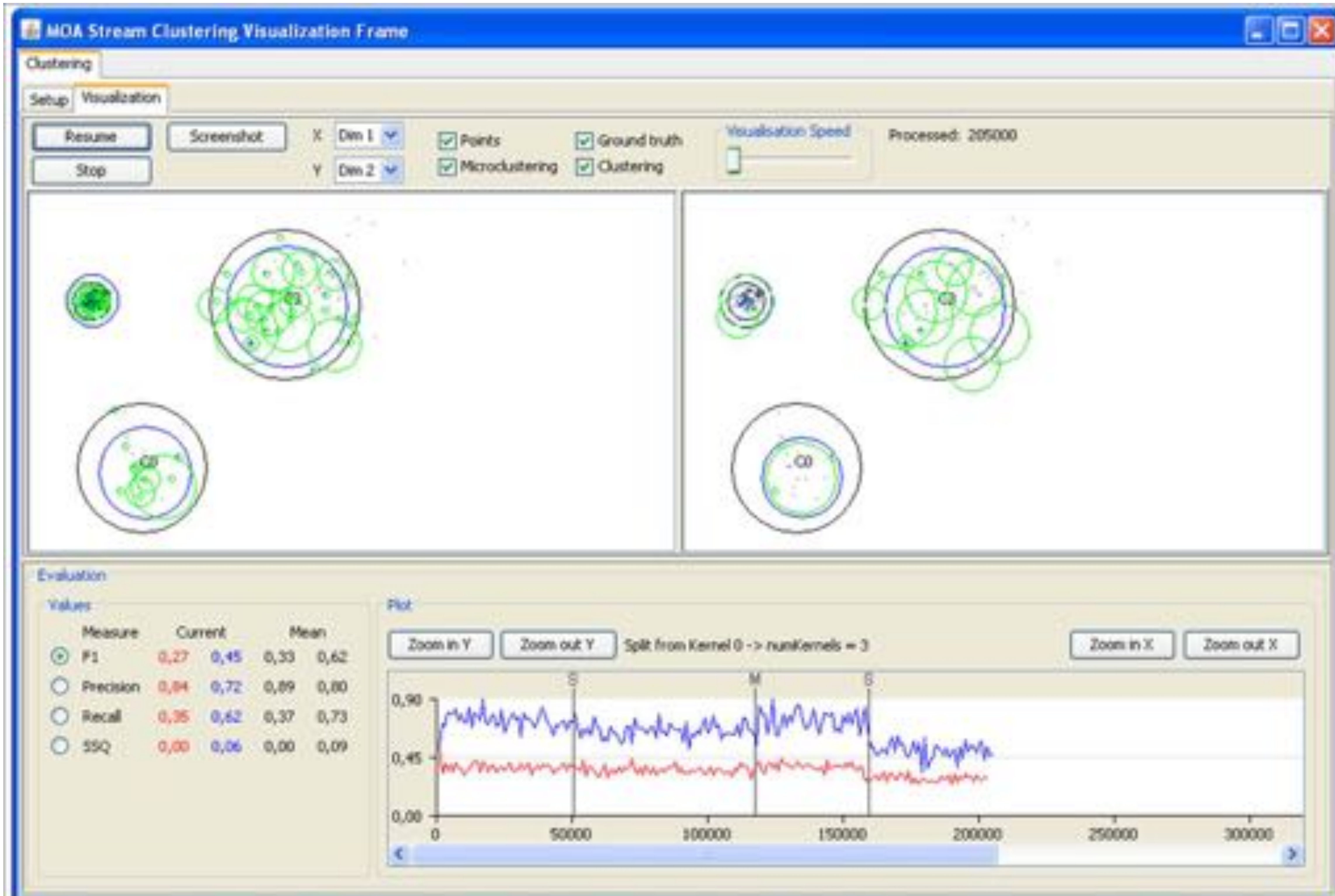
Prequential Evaluation

- The error of a model is computed from the sequence of examples.
- For each example in the stream, the actual model makes a prediction based only on the example attribute-values.



$$S = \sum_{i=1}^{n} L(y_i, \hat{y}_i).$$

Clustering



MOA Stream Clustering Visualization Frame
Clustering
Setup
Visualization
Resume
Screenshot
Stop
X Dim 1
Y Dim 2
Points
Microclustering
Ground truth
Clustering
Visualisation Speed
Processed: 205000
C0
C1
Evaluation
Values
Measure
Current
Mean
F1 0,27 0,45 0,33 0,62
Precision 0,84 0,72 0,89 0,80
Recall 0,35 0,62 0,37 0,73
SSQ 0,00 0,06 0,00 0,09
Plot
Zoom in Y
Zoom out Y
Split from Kernel 0 -> numKernels = 3
Zoom in X
Zoom out X
0,90
0,45
0,00
0
50000
100000
150000
200000
250000
300000

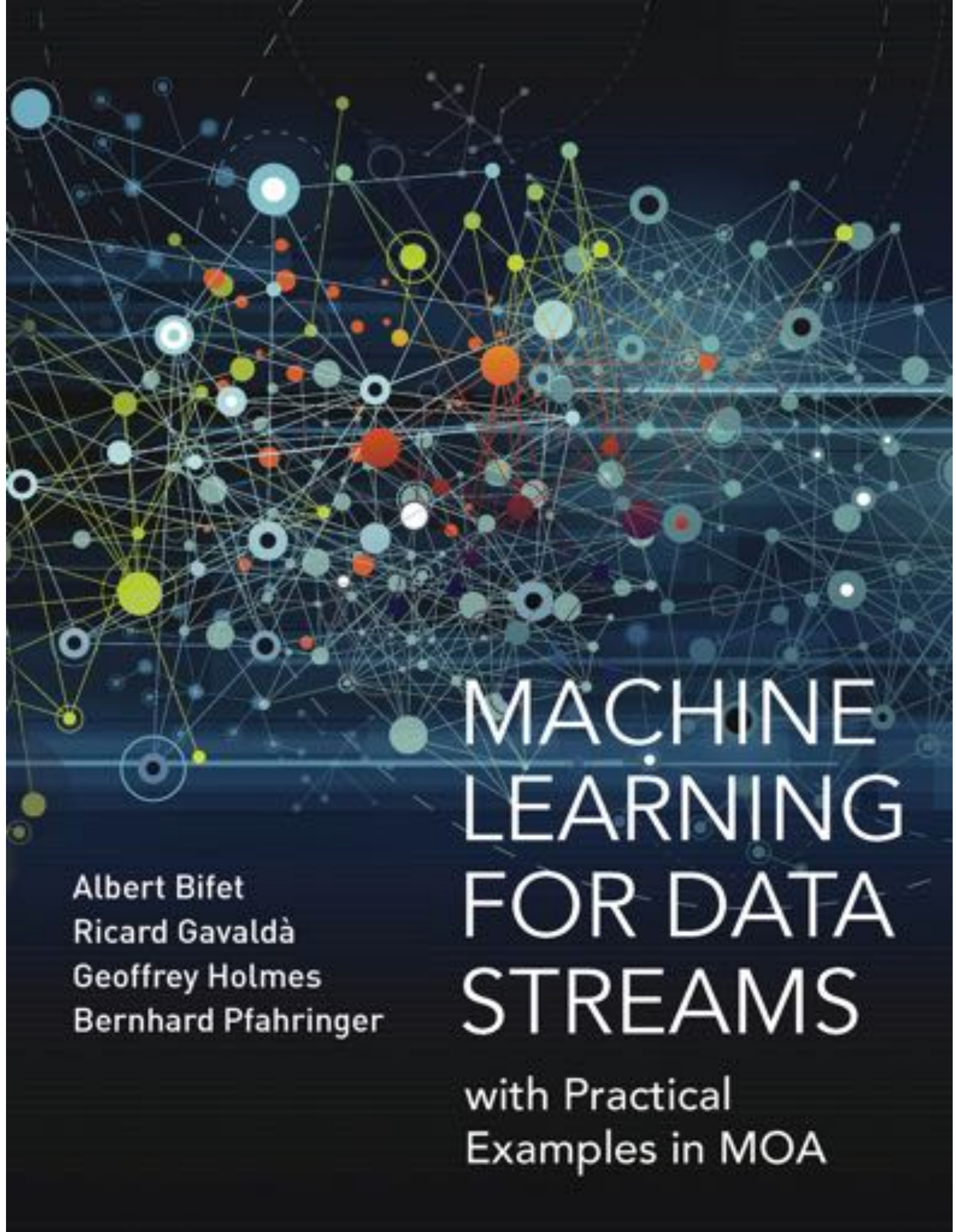
MOA Algorithms

- Multi-label/ Multi-target
- Outlier Detection
- Concept Drift Detection
- Active Learning
- Frequent Itemset Mining
- Frequent Graph Mining
- Recommendation Systems

Command Line

- `java -cp .:moa.jar:weka.jar -javaagent:sizeofag.jar moa.DoTask "EvaluatePeriodicHeldOutTest -l DecisionStump -s generators.WaveformGenerator -n 100000 -i 100000000 -f 1000000" > dsresult.csv`
- This command creates a comma separated values file:
  - training the DecisionStump classifier on the WaveformGenerator data,
  - using the first 100 thousand examples for testing,
  - training on a total of 100 million examples,
  - and testing every one million examples

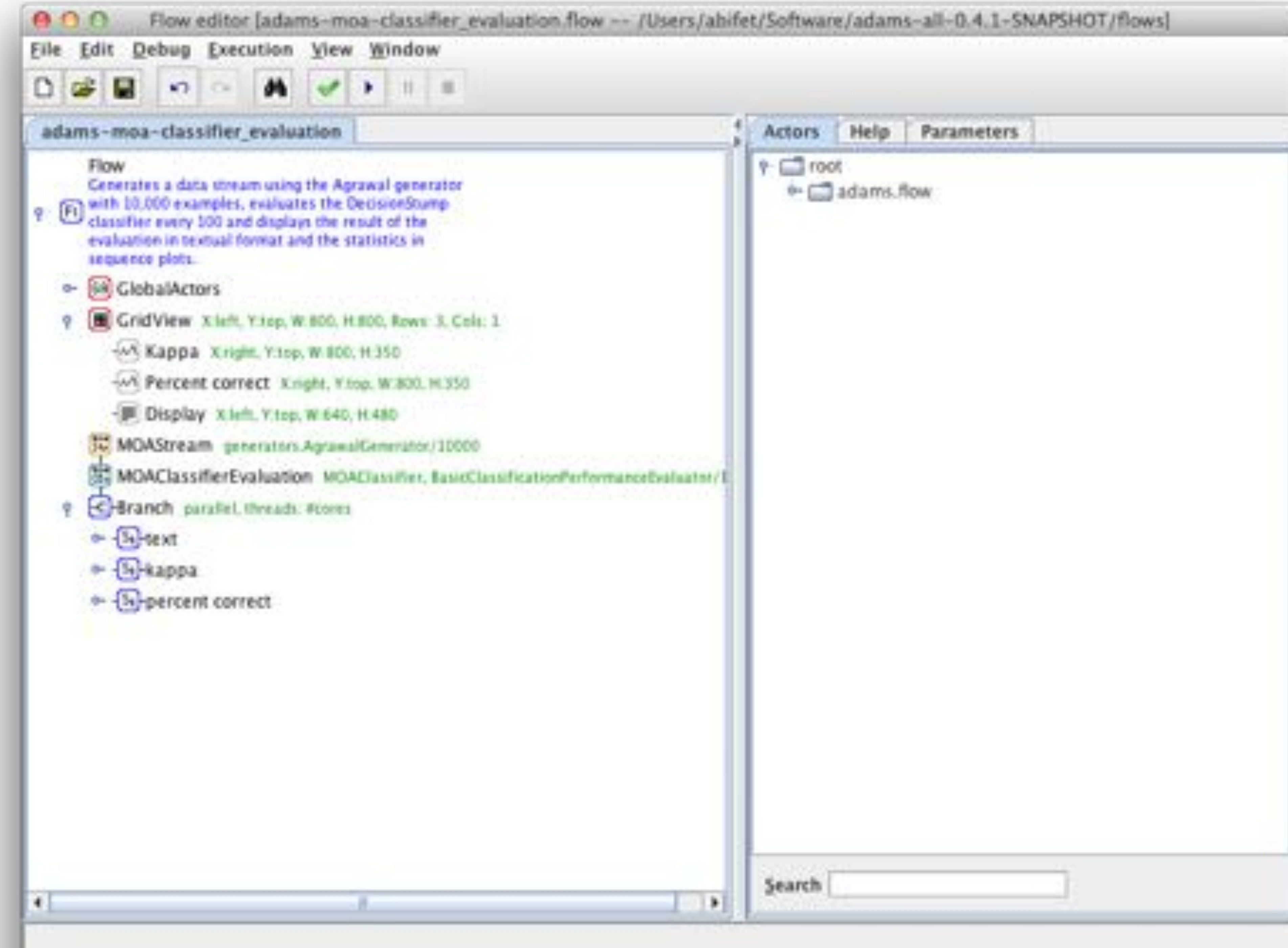
The MIT Press



ADAMS

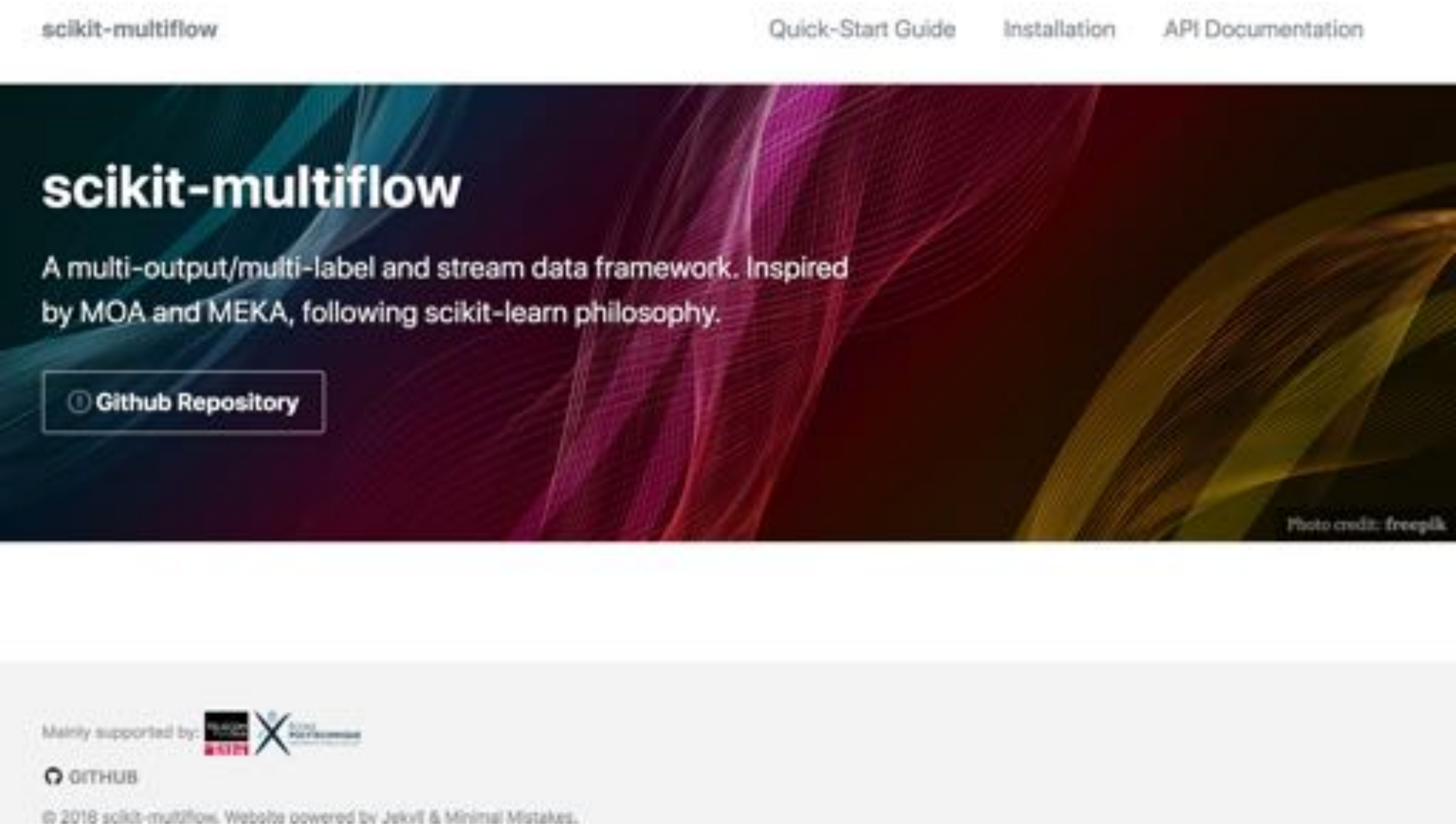
THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

Advanced Data Mining And Machine Learning System



OpenML

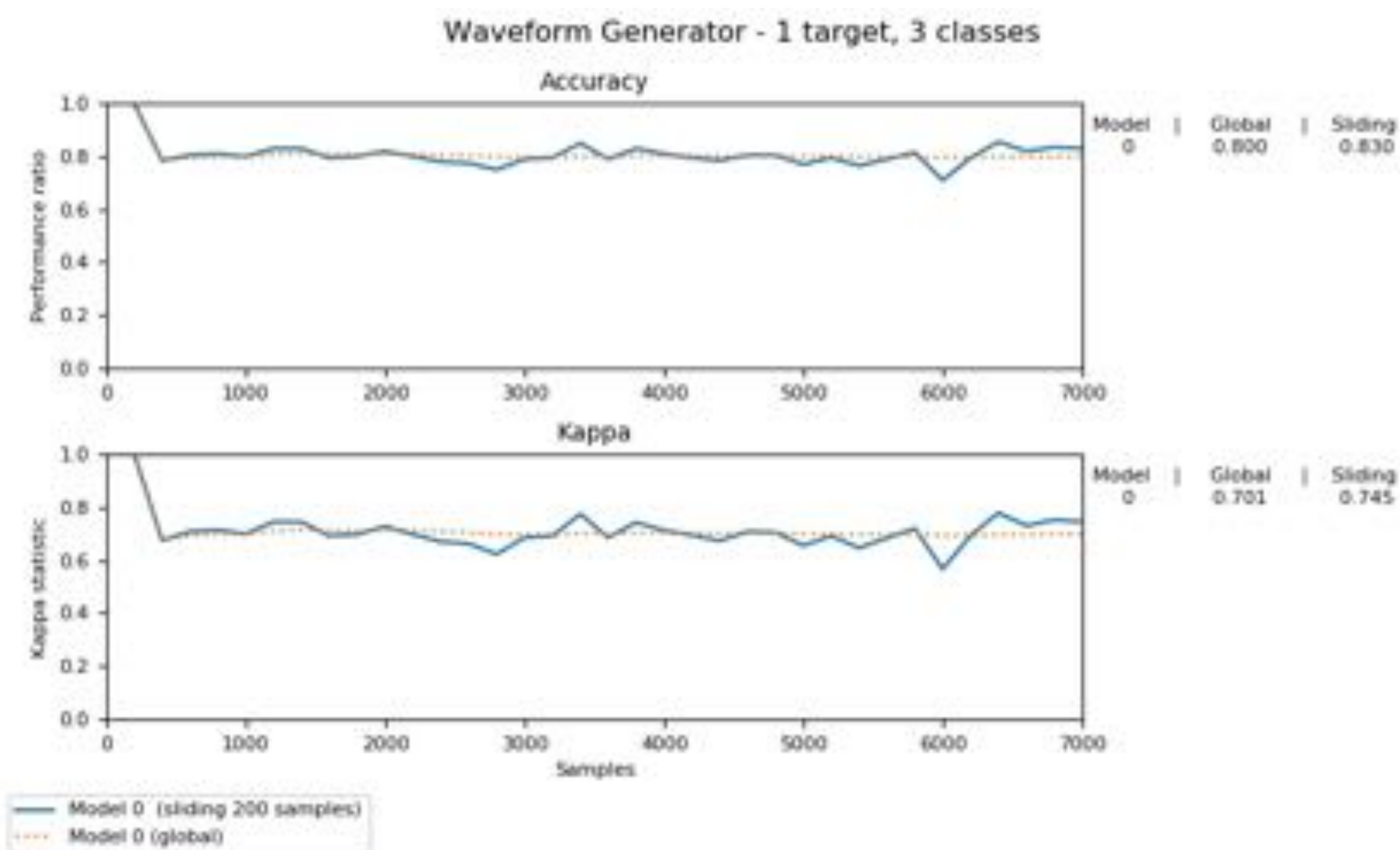
NETWORKED SCIENCE AND MACHINE LEARNING


scikit-multiflow
Quick-Start Guide
Installation
API Documentation
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:
GITHUB
© 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

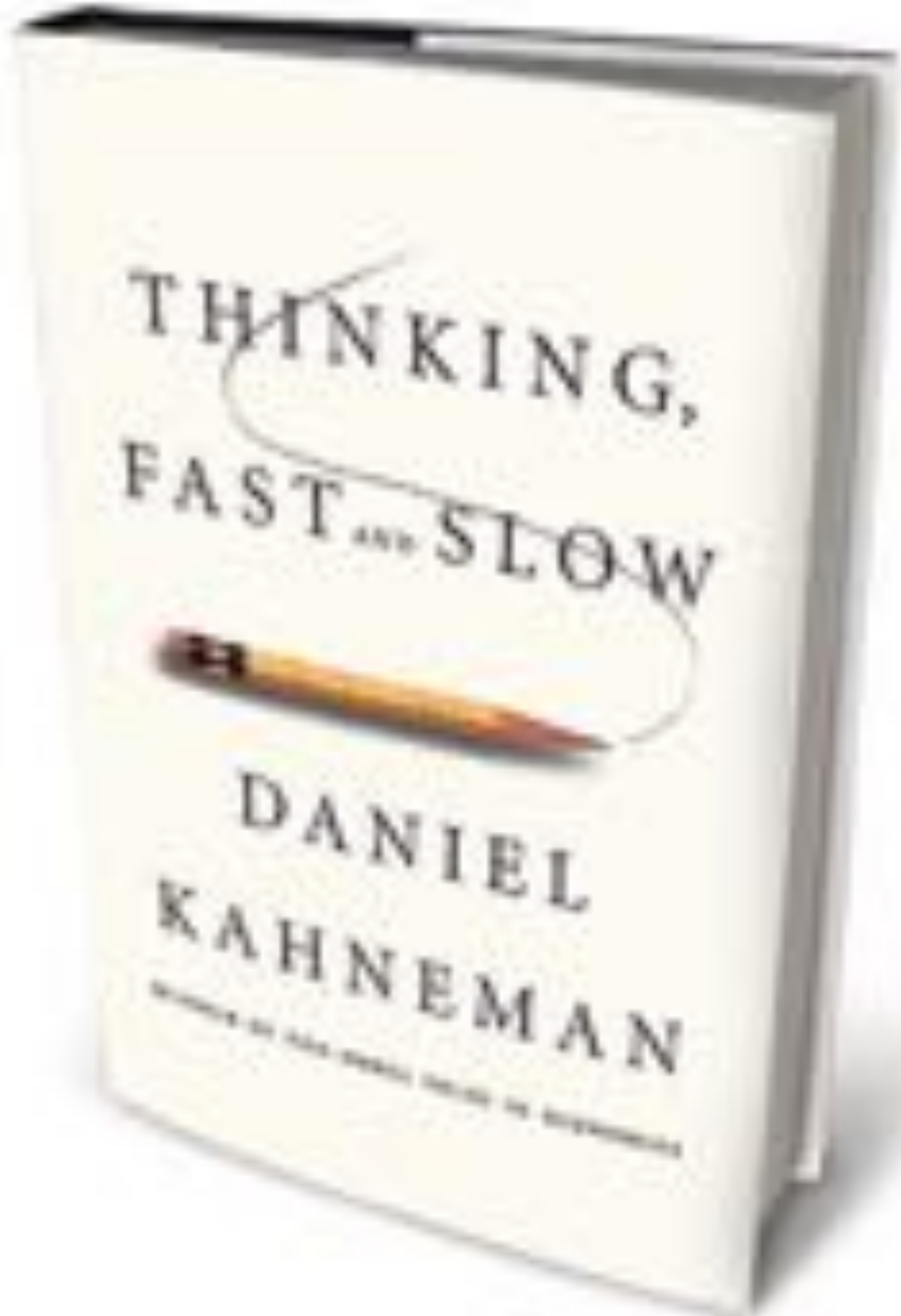


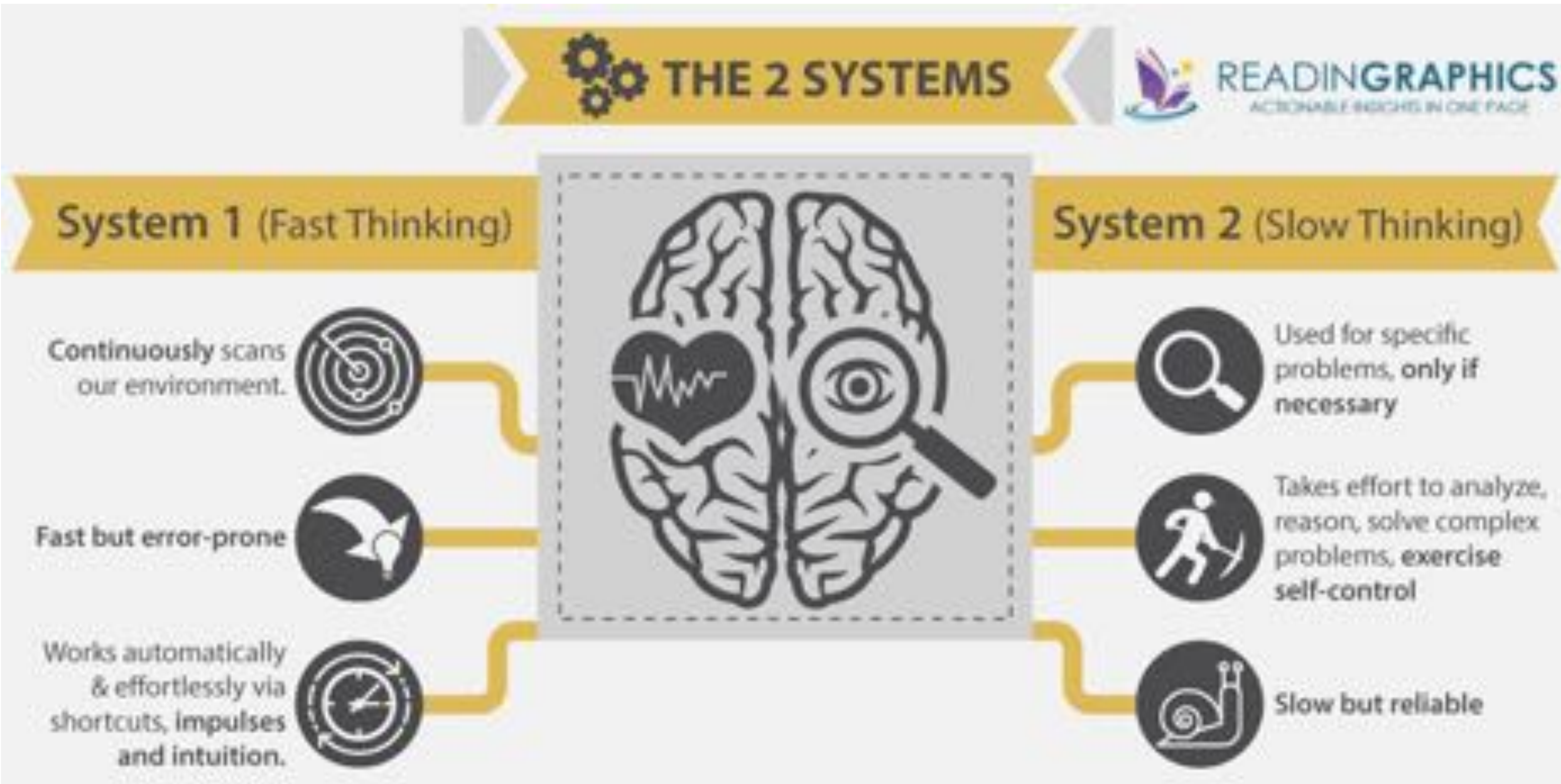
scikit-multiflow

Jesse Read
Ecole Polytechnique
France

Jacob Montiel
Telecom ParisTech
France

Learning Fast and Slow


THINKING,
FAST
SLOW
DANIEL
KAHNEMAN

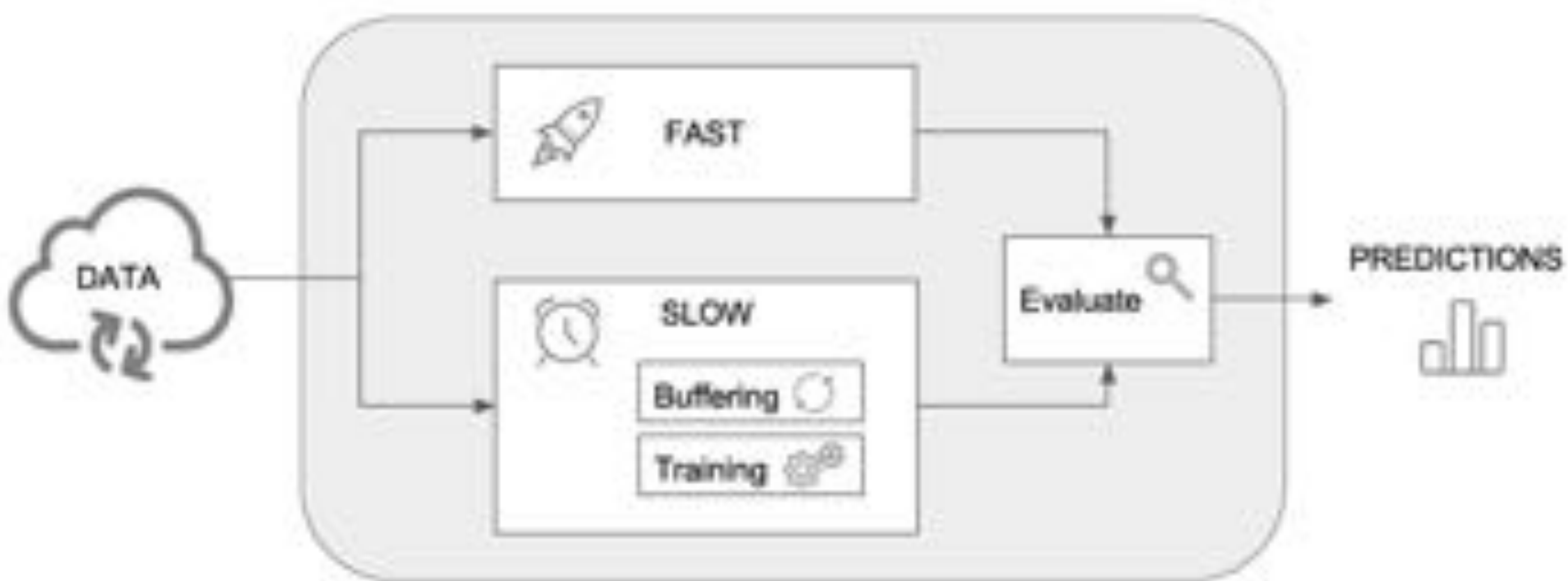

THE 2 SYSTEMS
READINGRAPHICS
System 1 (Fast Thinking)
System 2 (Slow Thinking)
Continuously scans our environment.
Fast but error-prone
Works automatically & effortlessly via shortcuts, impulses and intuition.
Used for specific problems, only if necessary
Takes effort to analyze, reason, solve complex 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

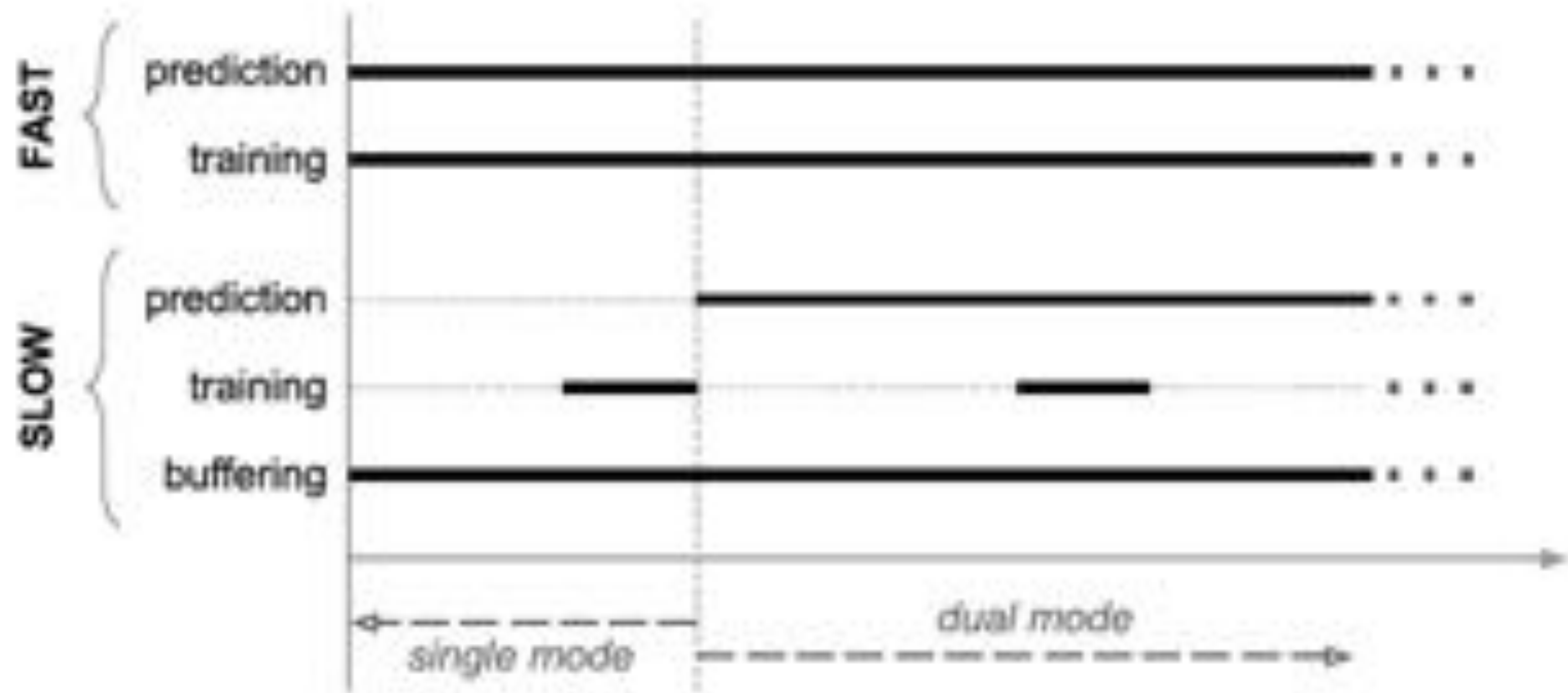


Figure 2: FSC operation modes.

Learning Fast and Slow

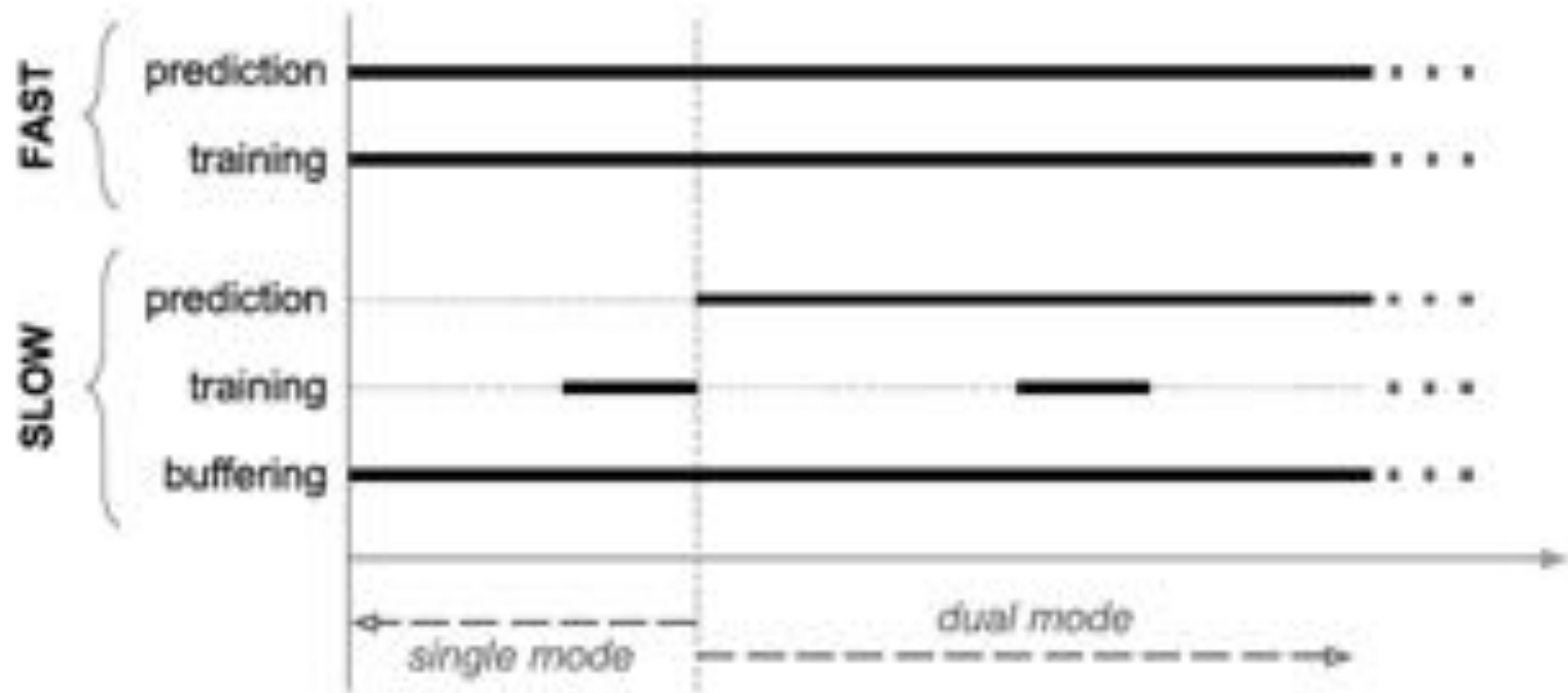
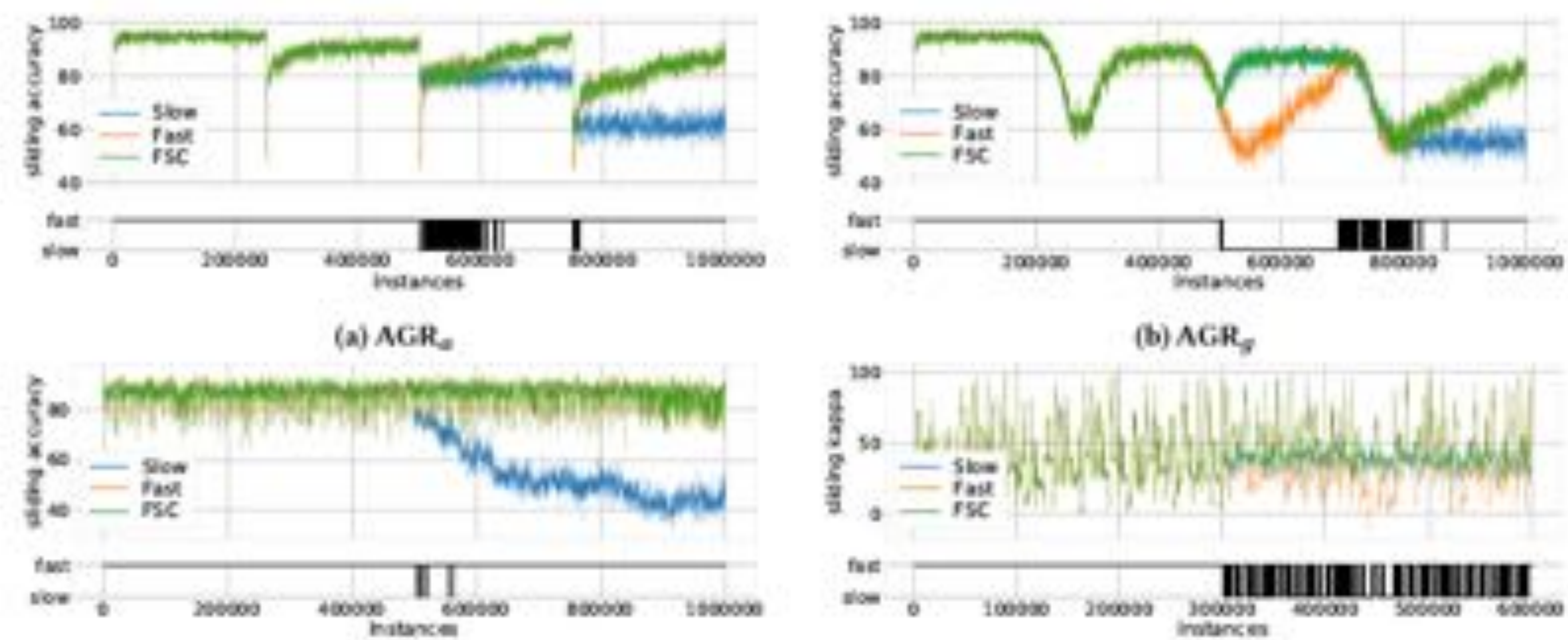


Figure 2: FSC operation modes.

Learning Fast and Slow



(a) AGR_a

(b) AGR_g

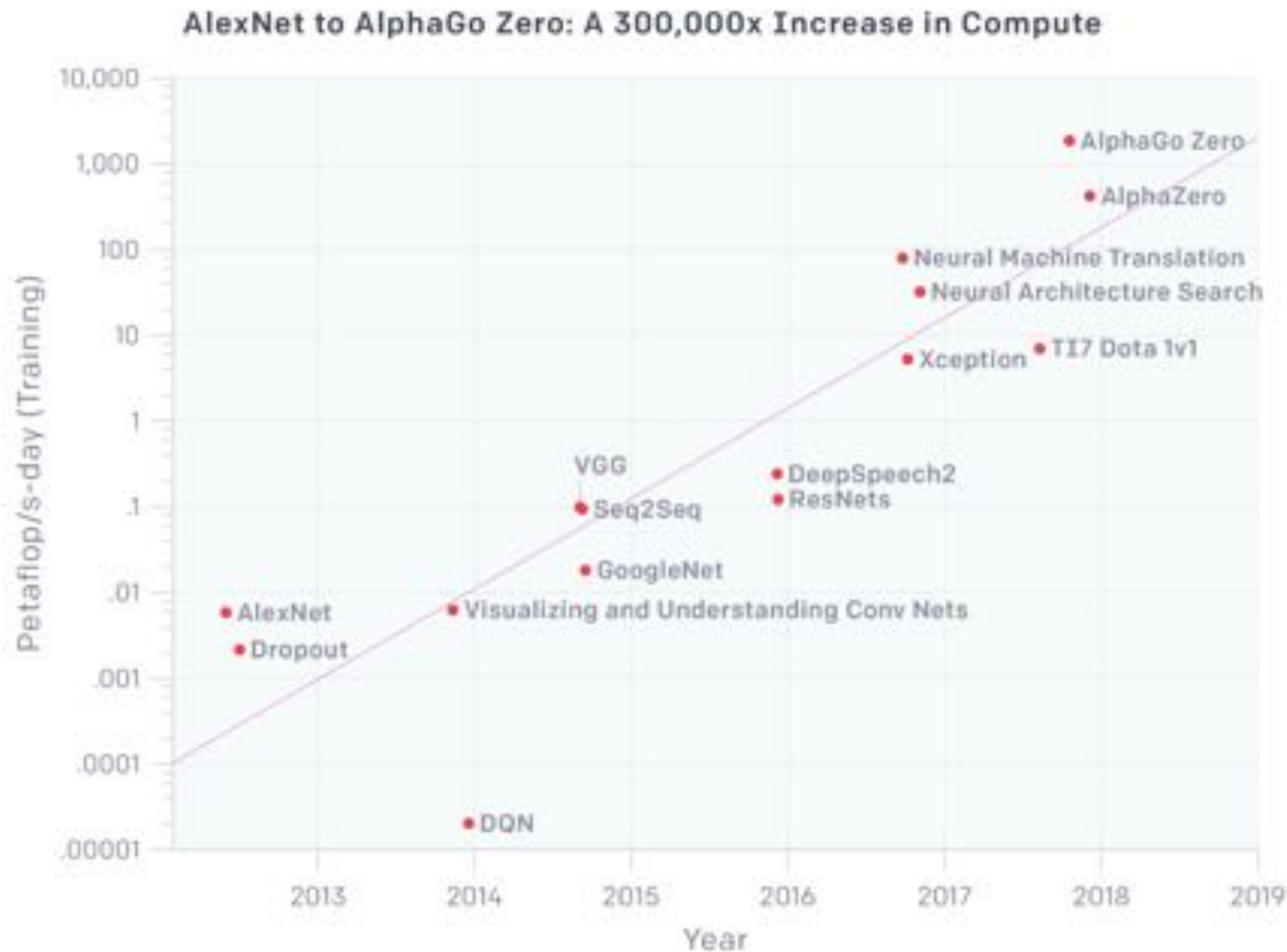
2. Green AI

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
Petaflop/s-day (Training)
10,000
1,000
100
10
1
.1
.01
.001
.0001
.00001
AlphaGo Zero
AlphaZero
Neural Machine Translation
Neural Architecture Search
TI7 Dota 1v1
Xception
VGG
Seq2Seq
DeepSpeech2
ResNets
GoogleNet
AlexNet
Dropout
Visualizing and Understanding Conv Nets
DQN
2013
2014
2015
2016
2017
2018
2019
Year

Creating an AI can be five times worse for the planet than a car

TECHNOLOGY 6 June 2019

Energy and Policy Considerations for Deep Learning in NLP

Emma Strubell **Ananya Ganesh** **Andrew McCallum**
College of Information and Computer Sciences
University of Massachusetts Amherst
{strubell, aganesh, mccallum}@cs.umass.edu

Abstract

Recent progress in hardware and methodology for training neural networks has ushered in a new generation of large networks trained on abundant data. These models have obtained notable gains in accuracy across many NLP tasks. However, these accuracy improvements depend on the availability of exceptionally large computational resources that necessitate similarly substantial energy consumption. As a result these models are costly to train and develop, both financially, due to the cost of hardware and electricity or cloud compute time, and environmentally, due to the carbon footprint required to fuel modern tensor processing hardware. In this paper we bring

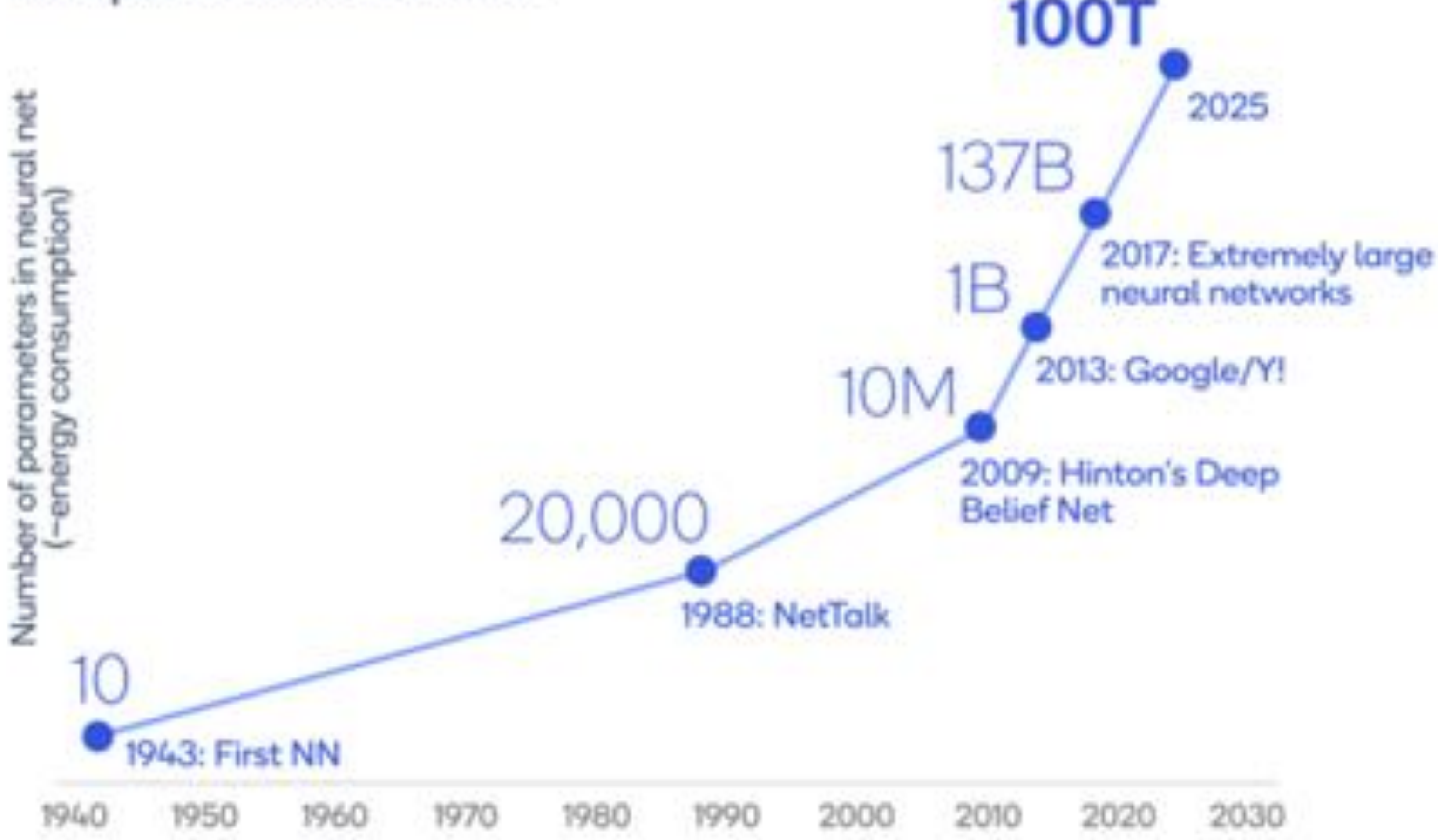
Consumption	**CO_2e (lbs)**
Air travel, 1 passenger, NY↔SF	1984
Human life, avg, 1 year	11,023
American life, avg, 1 year	36,156
Car, avg incl. fuel, 1 lifetime	126,000
Training one model (GPU)	
NLP pipeline (parsing, SRL)	39
w/ tuning & experimentation	78,468
Transformer (big)	192
w/ neural architecture search	626,155

Table 1: Estimated CO_2 emissions from training common NLP models, compared to familiar consumption.[1]

AI algorithms will be measured by the amount of intelligence they provide per joule.

Deep neural networks are energy hungry and growing fast

AI is being powered by the explosive growth of deep neural networks



Green AI

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

Approximation Algorithms

- What is the largest number that we can store in 8 bits?

1	0	1	0	1	0	1	0

Approximation Algorithms

- What is the largest number that we can store in 8 bits?

Programming Techniques

S.L. Graham, R.L. Rivest
Editors

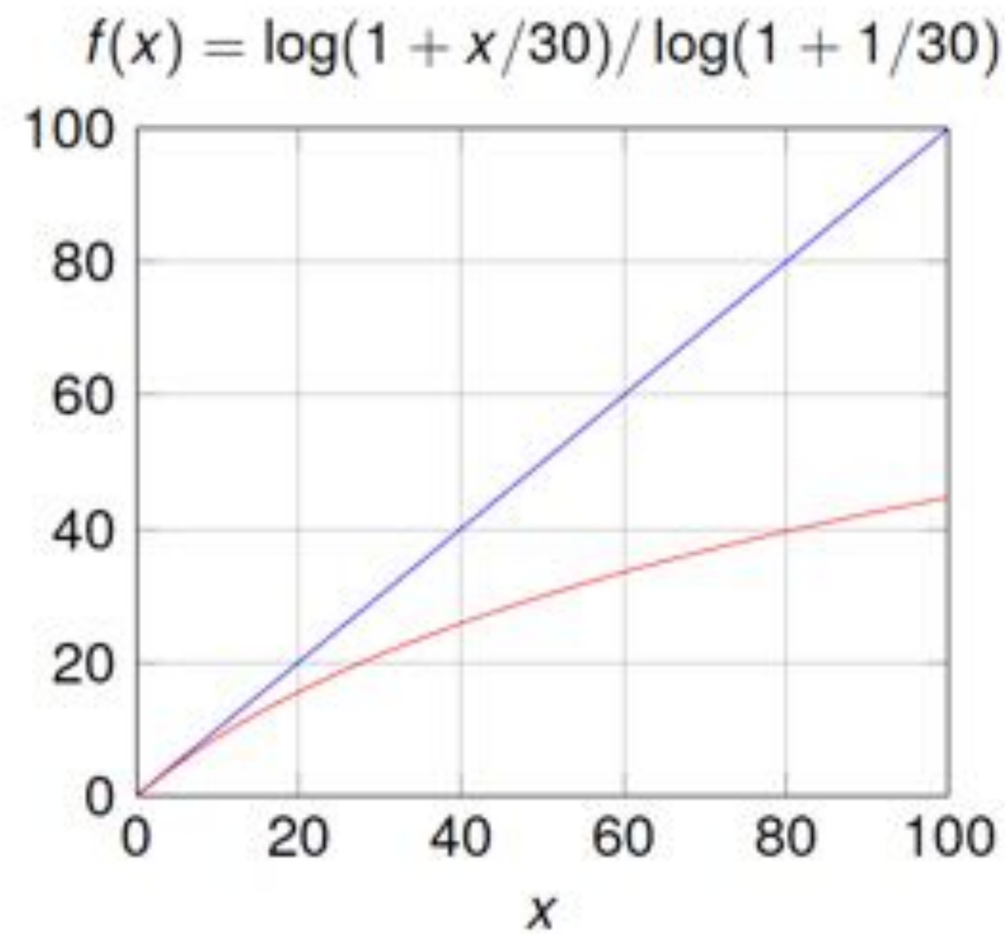
Counting Large Numbers of Events in Small Registers

Robert Morris
Bell Laboratories, Murray Hill, N.J.

It is possible to use a small counter to keep approximate counts of large numbers. The resulting expected error can be rather precisely controlled. An example is given in which 8-bit counters (bytes) are used to keep track of as many as 130,000 events with a relative error which is substantially independent of the number n of events. This relative error can be expected to be 24 percent or less 95 percent of the time (i.e. $\sigma = n/8$). The techniques could be used to advantage in multichannel counting hardware or software used for the monitoring of experiments or processes.

Approximation Algorithms

- What is the largest number that we can store in 8 bits?



$f(0) = 0, f(1) = 1$

Approximation Algorithms

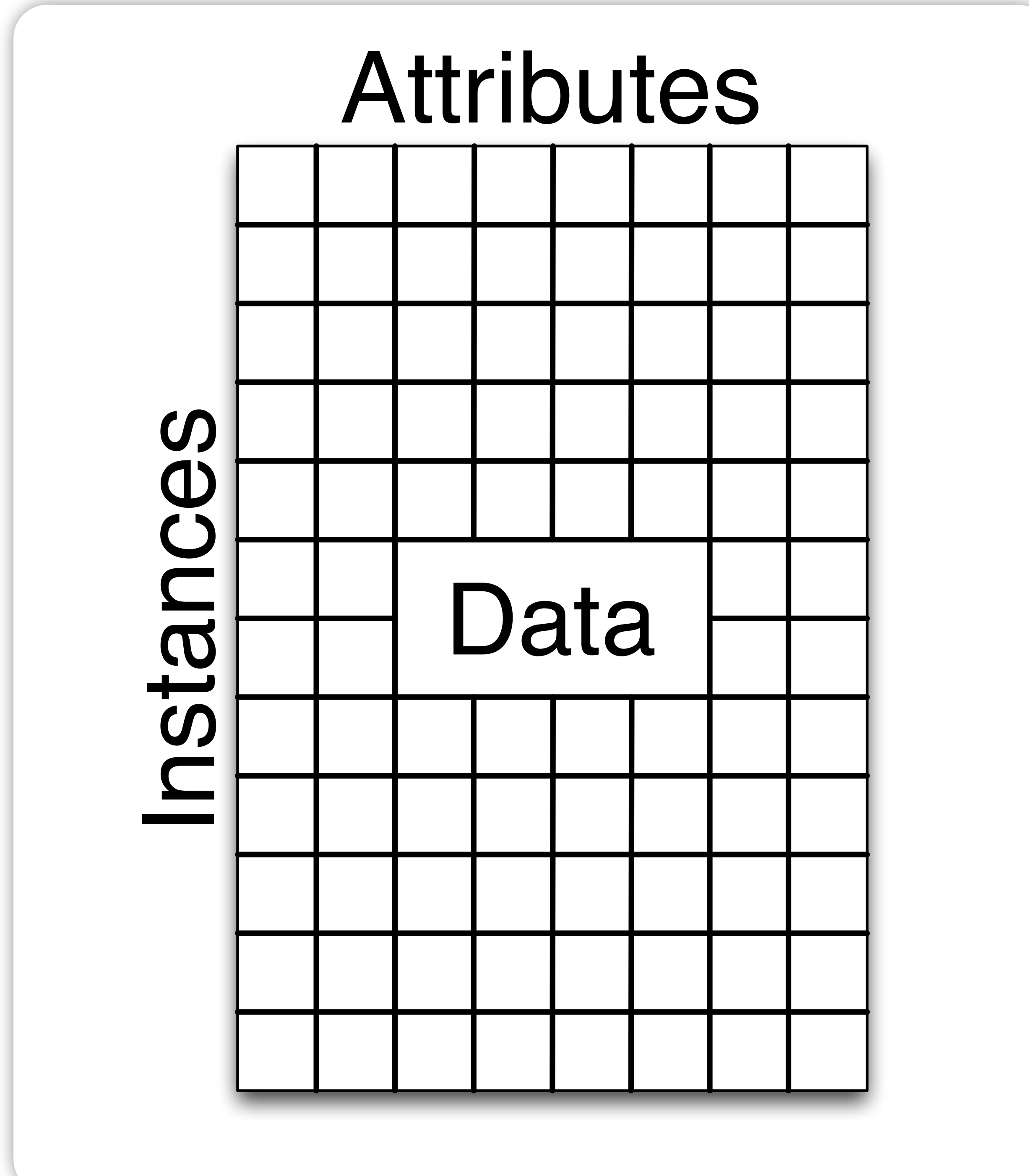
Morris Approximate Counting Algorithm

```
Init counter c ← 0
for every event in the stream
        do rand = random number between 0 and 1
           if rand < p
              then c ← c + 1
```

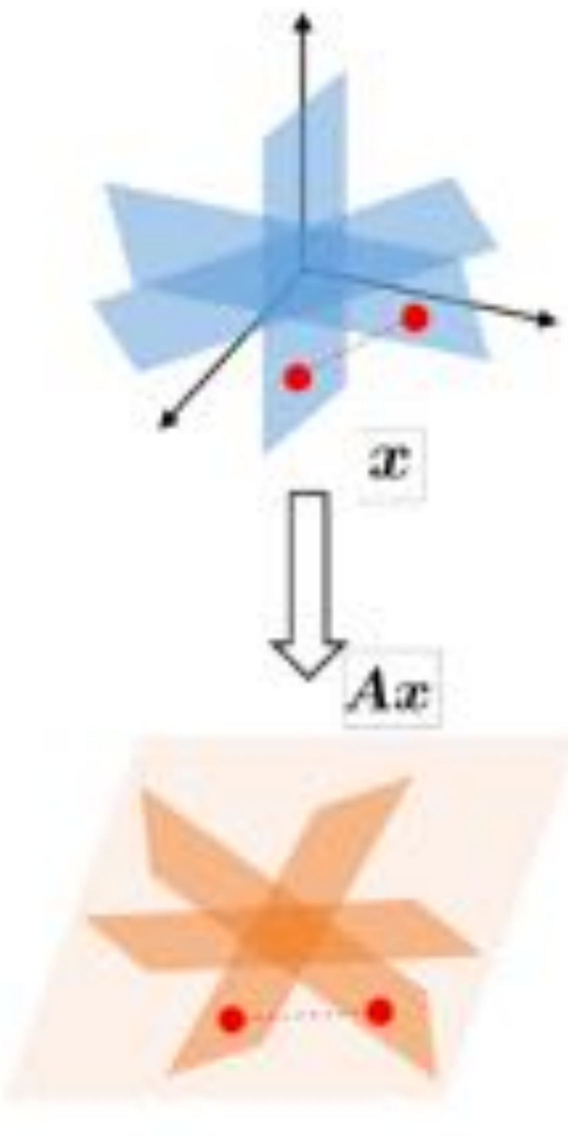
- What is the largest number that we can store in 8 bits?

Green AI

- Transform Big Data into Small Data
  - Vertical: reducing features
  - Horizontal: reducing instances
- Make data stream methods more energy efficient
  - Use Energy as a measure, not time and memory

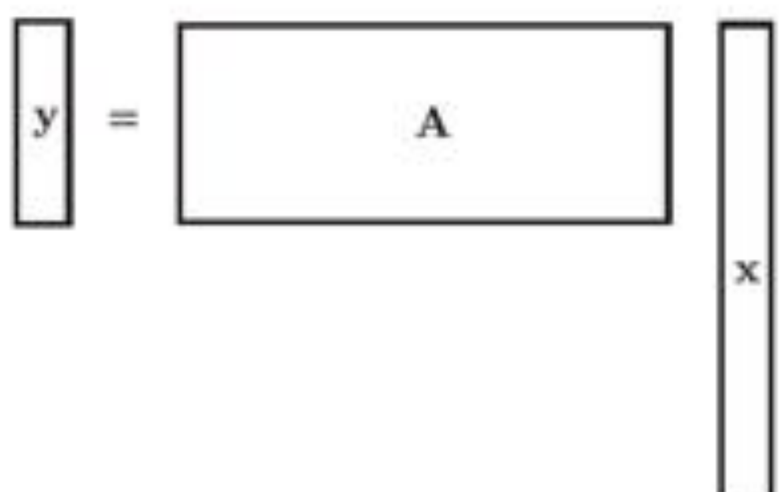


Compressed Sensing



Definition 1 (Restricted Isometry Property) *A $m \times n$ sensing matrix A satisfies the restricted isometry property, (k, ϵ)-RIP, if it acts as a near-isometry with distortion factor ϵ, over all k-sparse vectors. In other words, for any k-sparse vector $\boldsymbol{x} \in \mathbb{R}^n$ the following near-isometry property holds:*

$$(1 - \epsilon)\|\boldsymbol{x}\|_2 \leq \|A\boldsymbol{x}\|_2 \leq (1 + \epsilon)\|\boldsymbol{x}\|_2.$$



Joint Work with:
- **Maroua Bahri**
- **Silviu Maniu**
- **Nikos Tziortziotis**
- **Rodrigo Mello**

Coresets

Coreset of a set P with respect to some problem

Small subset that approximates the original set P.

- Solving the problem for the coreset provides an approximate solution for the problem on P.

(k, ϵ)-coreset

A (k, ϵ)-coreset S of P is a subset of P that for each C of size k

$$(1 - \epsilon)cost(P, C) \leq cost_w(S, C) \leq (1 + \epsilon)cost(P, C)$$

3. 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

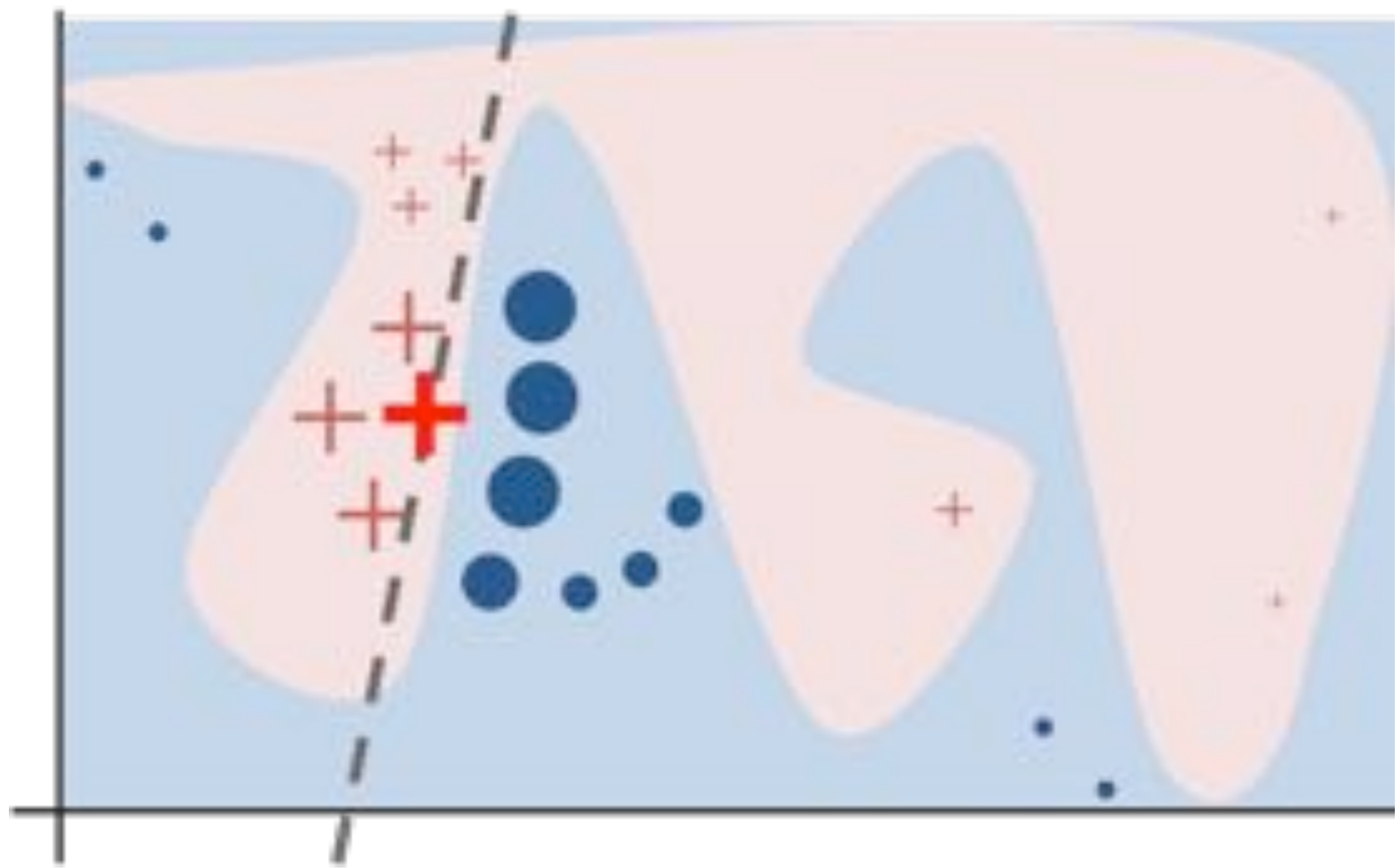
343 248 people are talking about this

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.

Lime: Explaining the predictions of any machine learning classifier



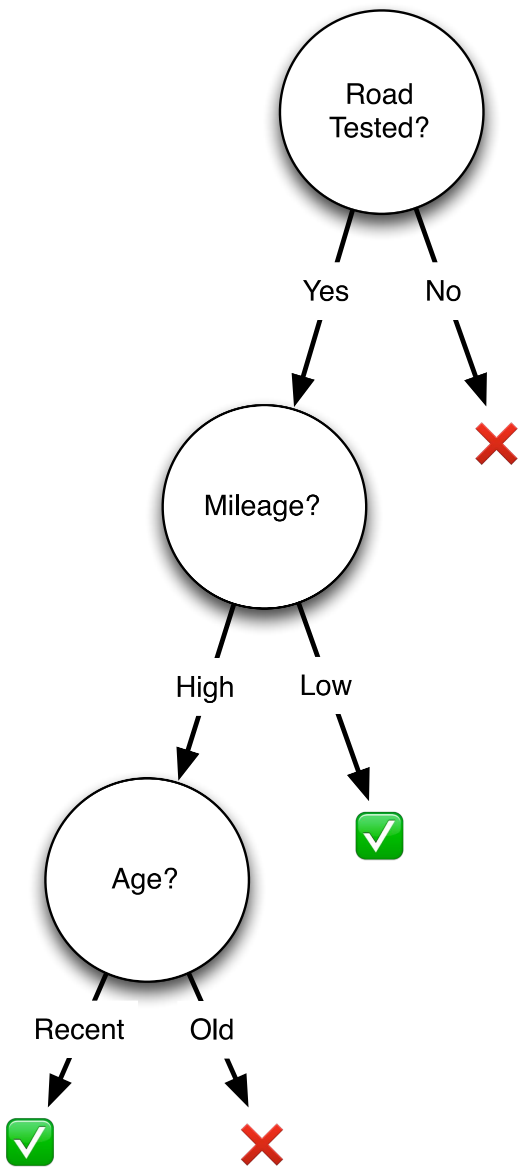
"Why Should I Trust You?": Explaining the Predictions of Any Classifier

Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin, KDD 2016

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 $G(x_1) - G(x_2) > \varepsilon = \sqrt{\frac{R^2 \ln(1/\delta)}{2n}}$

TensorForest: Scalable Random Forests on TensorFlow

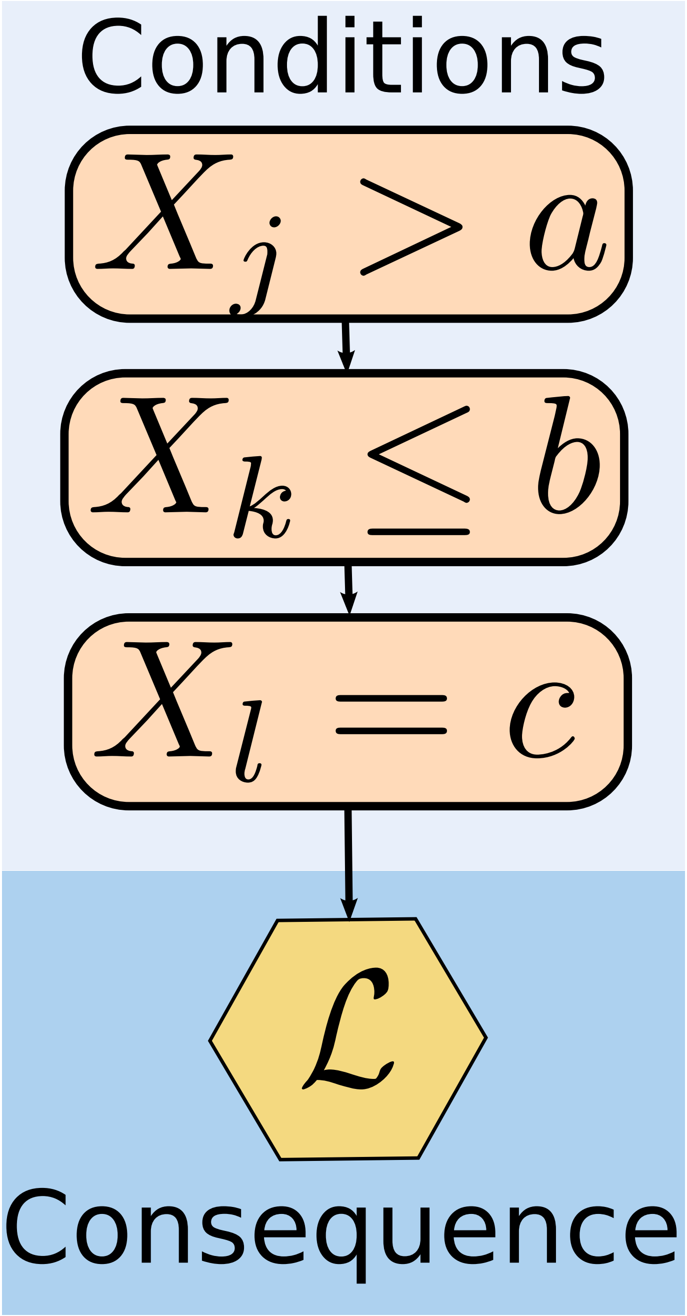
Thomas Colthurst, Gilbert Hendry, Zachary Nado, D. Sculley
Google Inc.
{thomaswc, gilberth, znado, dsculley}@google.com

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

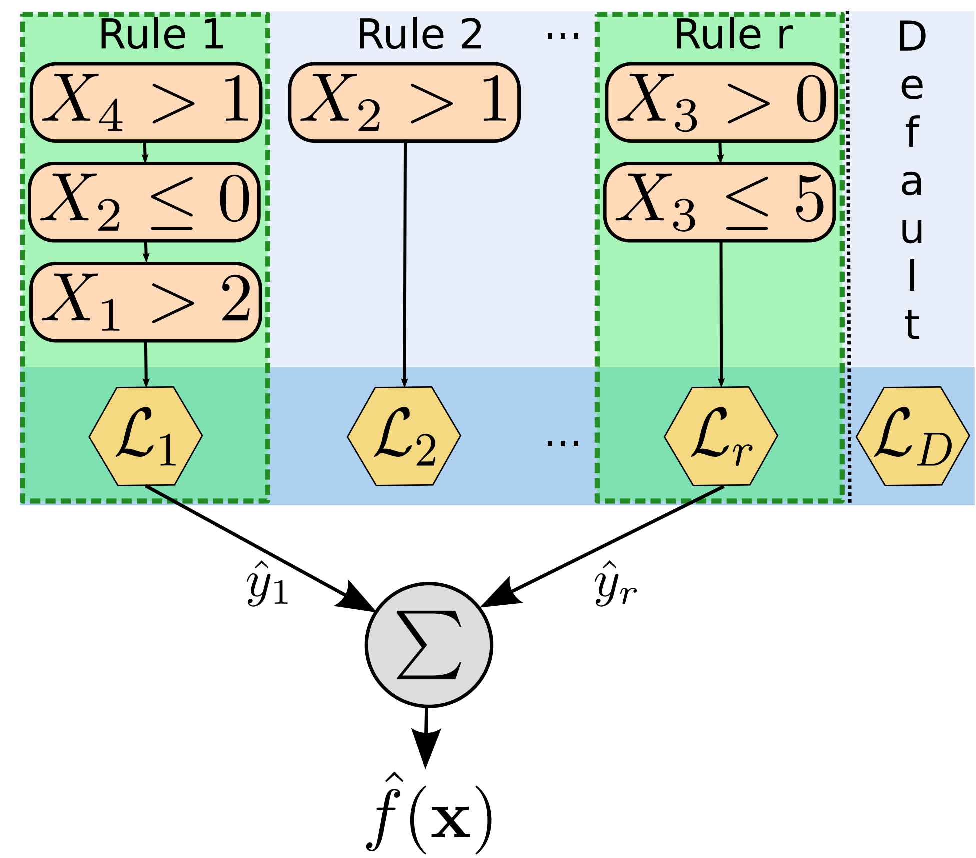
- Problem: very large decision trees have context that is complex and hard to understand
- Rules: self-contained, modular, easier to interpret, no need to cover universe
- $\mathcal{L}$ keeps sufficient statistics to:
  - make predictions
  - expand the rule
  - detect changes and anomalies



Adaptive Model Rules

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

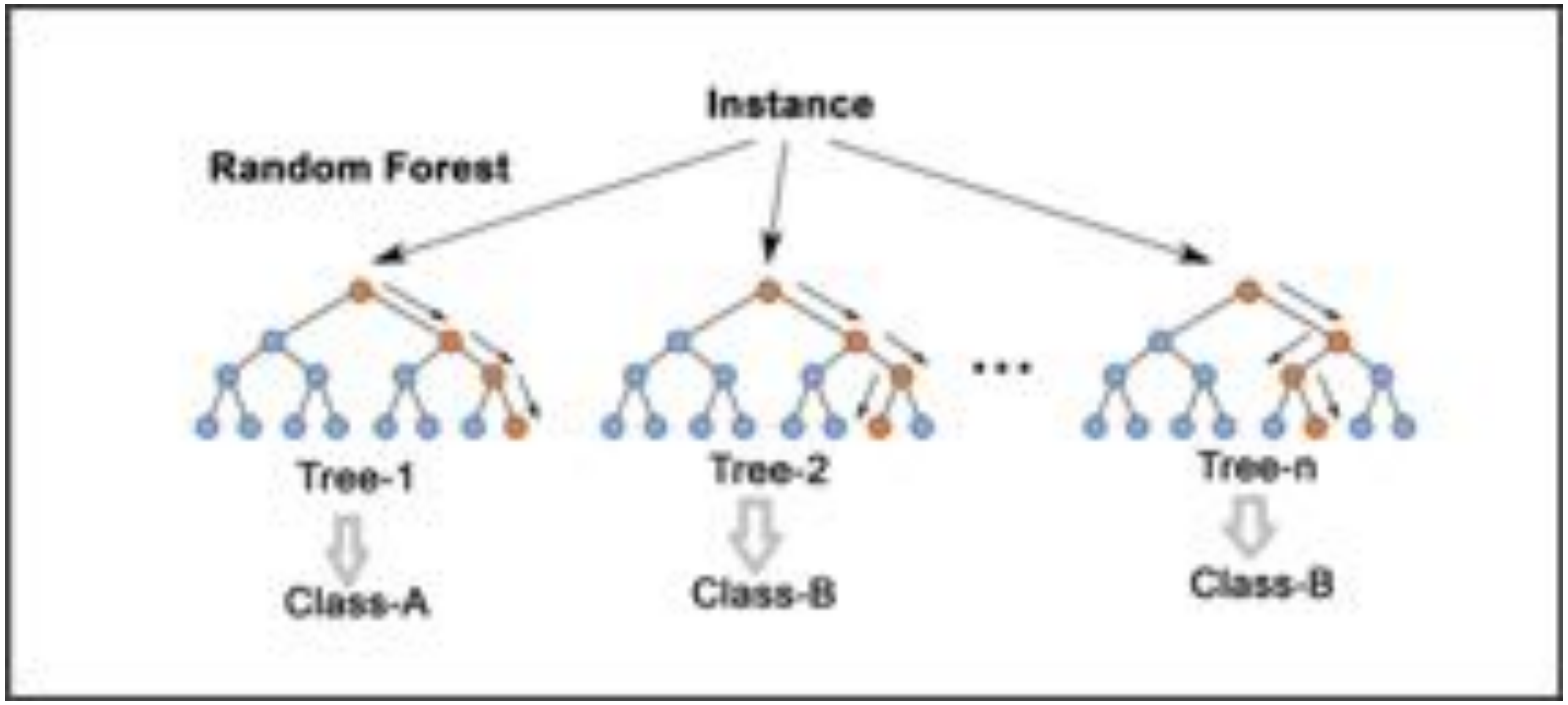
- Ruleset: ensemble of rules
- Rule prediction: mean, linear model
- Ruleset prediction
  - Weighted avg. of predictions of rules covering instance x
  - Weights inversely proportional to error
  - Default rule covers uncovered instances



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



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

ADWIN

ADWIN

An adaptive sliding window whose size is recomputed online according to the rate of change observed.

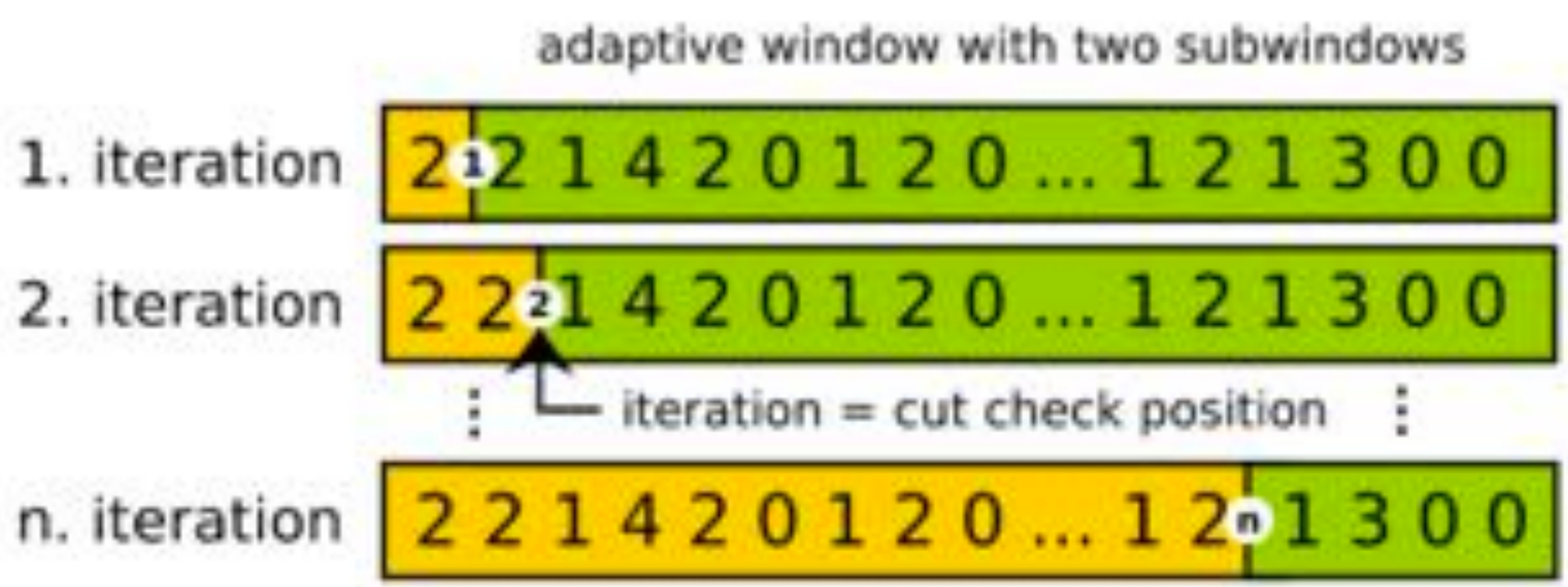


Figure 1: Iterations of the cut check procedure in ADWIN

ADWIN

- Classification
  - Adaptive Naive Bayes (Bifet et al. 2007)
  - Decision Trees: Hoeffding Adaptive Trees (Bifet et al. 2009)
  - `ADWIN` Bagging (Bifet et al. 2009)
  - Leveraging Bagging (Bifet et al. 2010)
  - Stacking of Restricted Hoeffding Trees (Bifet et al. 2012)
  - Multilabel Classification (Read et al. 2012)
  - Adaptive kNN (Bifet et al. 2013)
  - Random Forests (Marron et al. 2014)
- Frequent Pattern Mining
  - Frequent Closed Tree Mining (Bifet et al. 2008)
  - Frequent Closed Graph Mining (Bifet et al. 2011)

4. 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.

Art. 17 GDPR

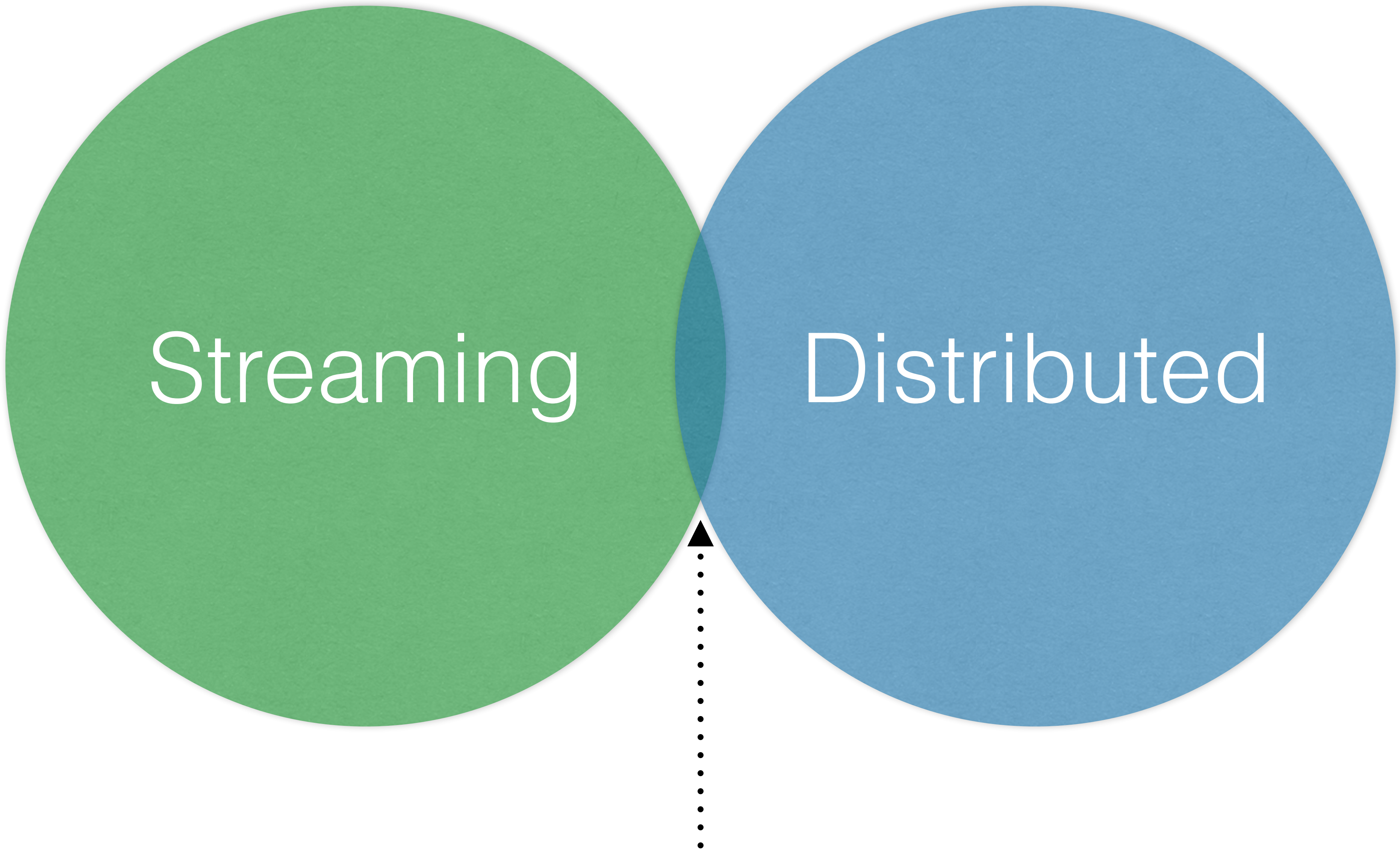
Right to erasure ('right to be forgotten')

(1) The data subject shall have the right to obtain from the controller the erasure of personal data concerning him or her without undue delay and the controller shall have the obligation to erase personal data without undue delay where one of the following grounds applies:

Should data have an expiration date?

5. Distributed Machine Learning for Data Streams

Vision

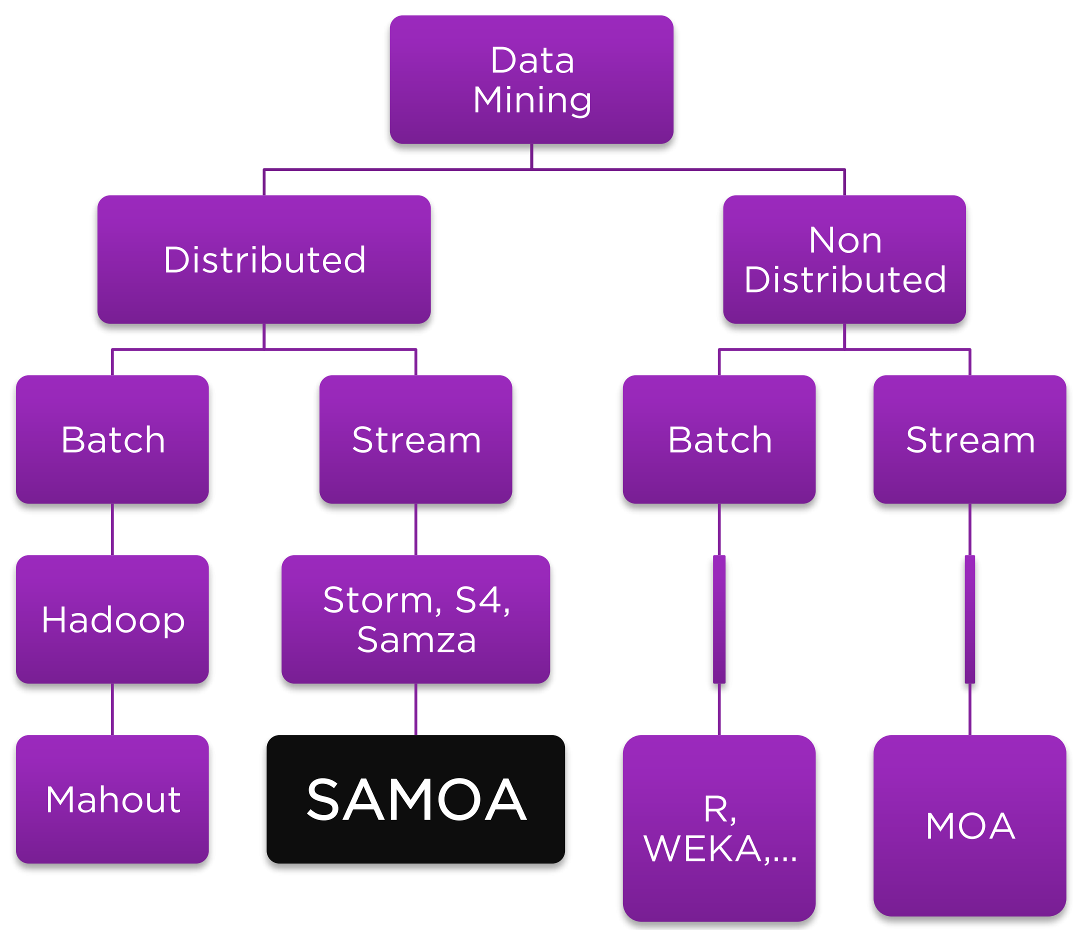


IoT Big Data Stream Mining

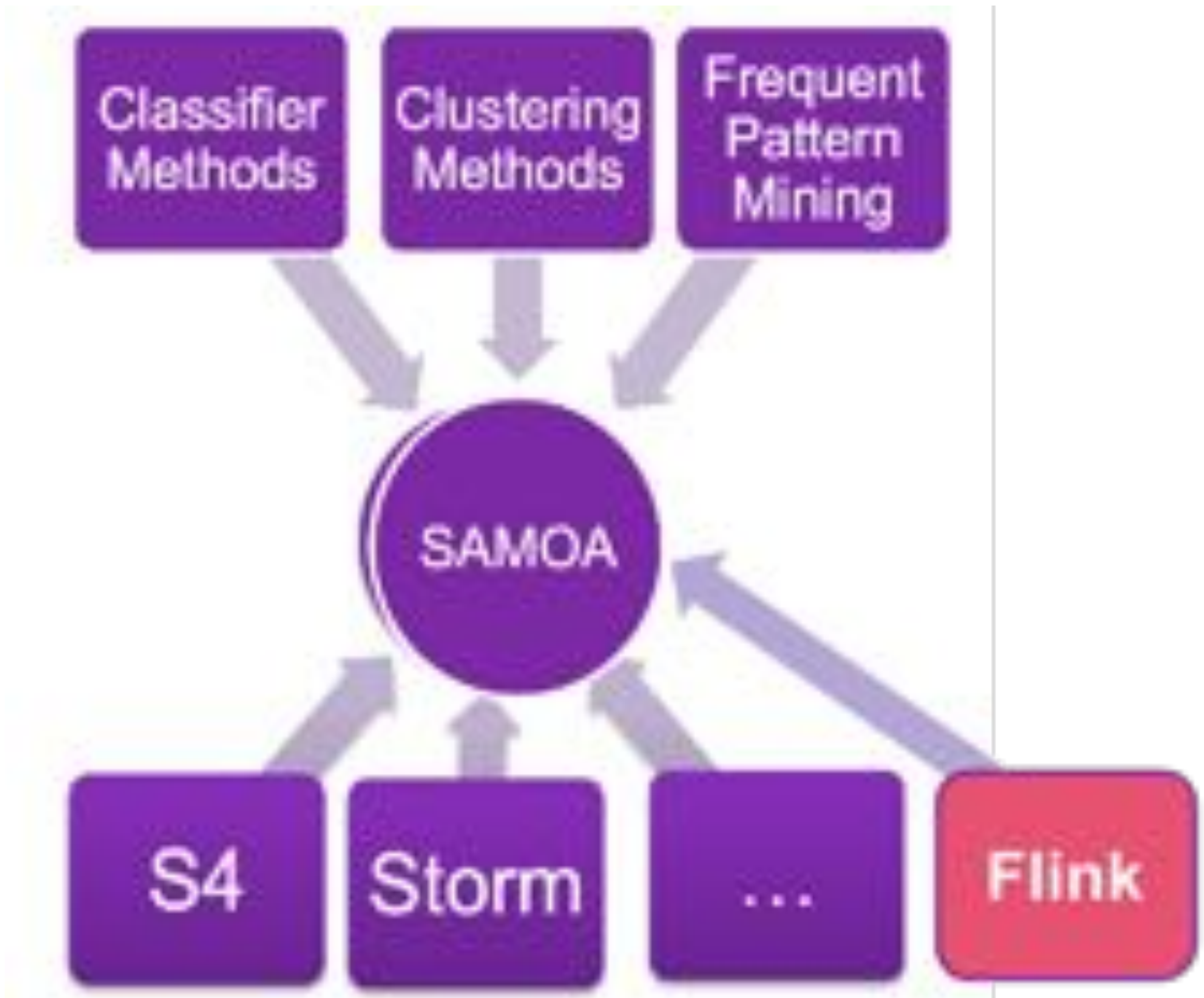
http://samoa-project.net

APACHE SAMOA

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

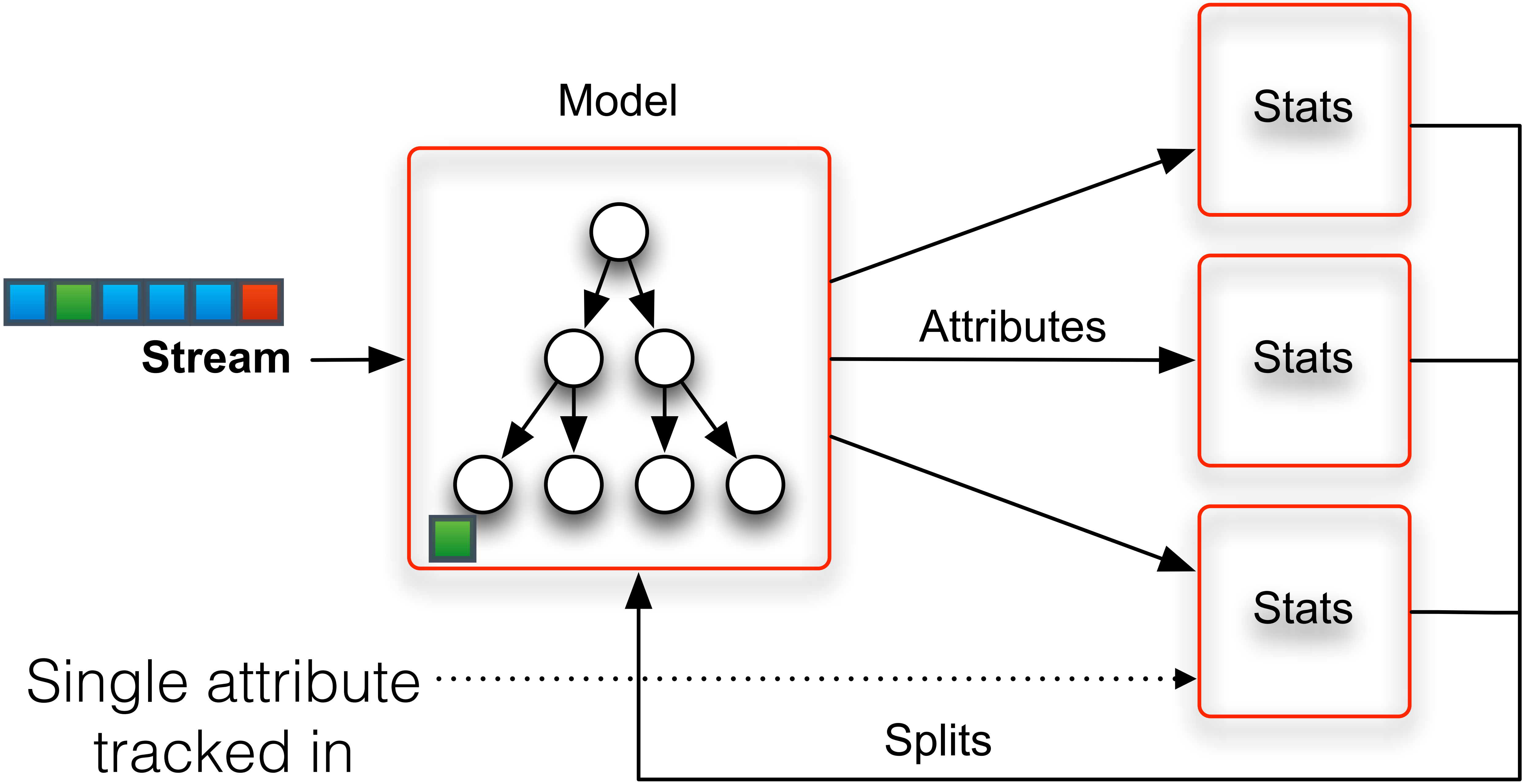


SAMOA ARCHITECTURE



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

Apache SAMOA Team

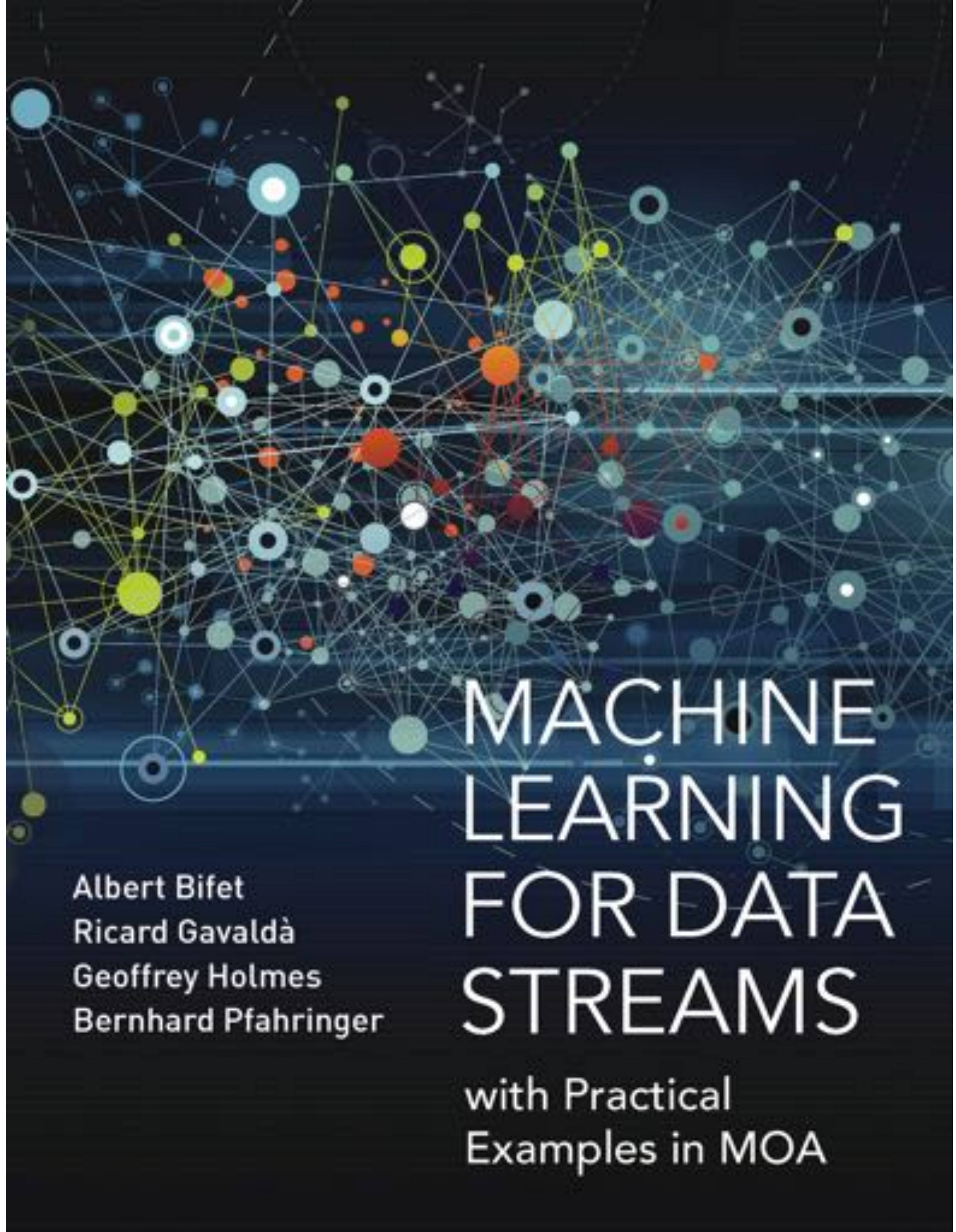
Gianmarco De Francisci Morales, Nicolas Kourtellis, Matthieu Morel, Arinto Murdopo, Antonio Severien, and Olivier Van Laere http://huawei-noah/github.io/streamDM

StreamDM

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

The MIT Press



Green Data Mining

GREEN
DATA MINING

Home CFP Dates Submission Organization

Second International Workshop on Energy Efficient Scalable Data Mining and Machine Learning

Co-located with ECML PKDD 2019
September 16, 2019 - Würzburg, Germany

Thanks!

@abifet

Machine Learning for Data Streams

Albert Bifet (@abifet)

Paris
TMA Conference 2019
20 June 2019

TELECOM Paris
IP PARIS
INSTITUT POLYTECHNIQUE DE PARIS
THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato

The MIT Press