Institut Mines-Télécom

Futur & Ruptures a 10 ANS ! 2008-2018

FONDATION Mines-Télécom La Fondation de l'IMT

INSTITUT CARNOT Télécom & Société numérique

ARTIFICIAL INTELLIGENCE AND DATA STREAM MINING

ALBERT BIFET

15 FEVRIER 2018

#Futur&Ruptures

IMT Atlantique Bretagne-Pays de la Loire École Mines-Télécom · IMT Lille Douai École Mines-Télécom IMT-Université de Lille · IMT Mines Albi-Carmaux École Mines-Télécom · IMT Mines Alès École Mines-Télécom · MINES Saint-Étienne Une école de l'IMT · TELECOM École de Management Une école de l'IMT · TELECOM ParisTech Une école de l'IMT · TELECOM SudParis Une école de l'IMT · EURECOM · MINES ParisTech · MINES nancy · ARMINES

SOMMAIRE

Institut Mines-Télécom

ARTIFICIAL INTELLIGENCE

Institut Mines-Télécom

Artificial Intelligence is the new Electricity

- Big Data
- **Internet of Things**
- Data Science
- Artificial Intelligence

What is AI?

- Artificial intelligence (AI) is an area of **computer science** that emphasizes the creation of intelligent machines".

ARTIFICIAL INTELLIGENCE CHALLENGES

Institut Mines-Télécom

Artificial Intelligence

HUFFPOST
UNITED KINGDOM

Edition: UK | Search HuffPost | Like 153k | Follow

FRONT PAGE NEWS POLITICS LIFESTYLE PARENTS TECH YOUNG VOICES COMEDY ENTERTAINMENT STYLE WHAT'S WORKING BLOGS

TECH

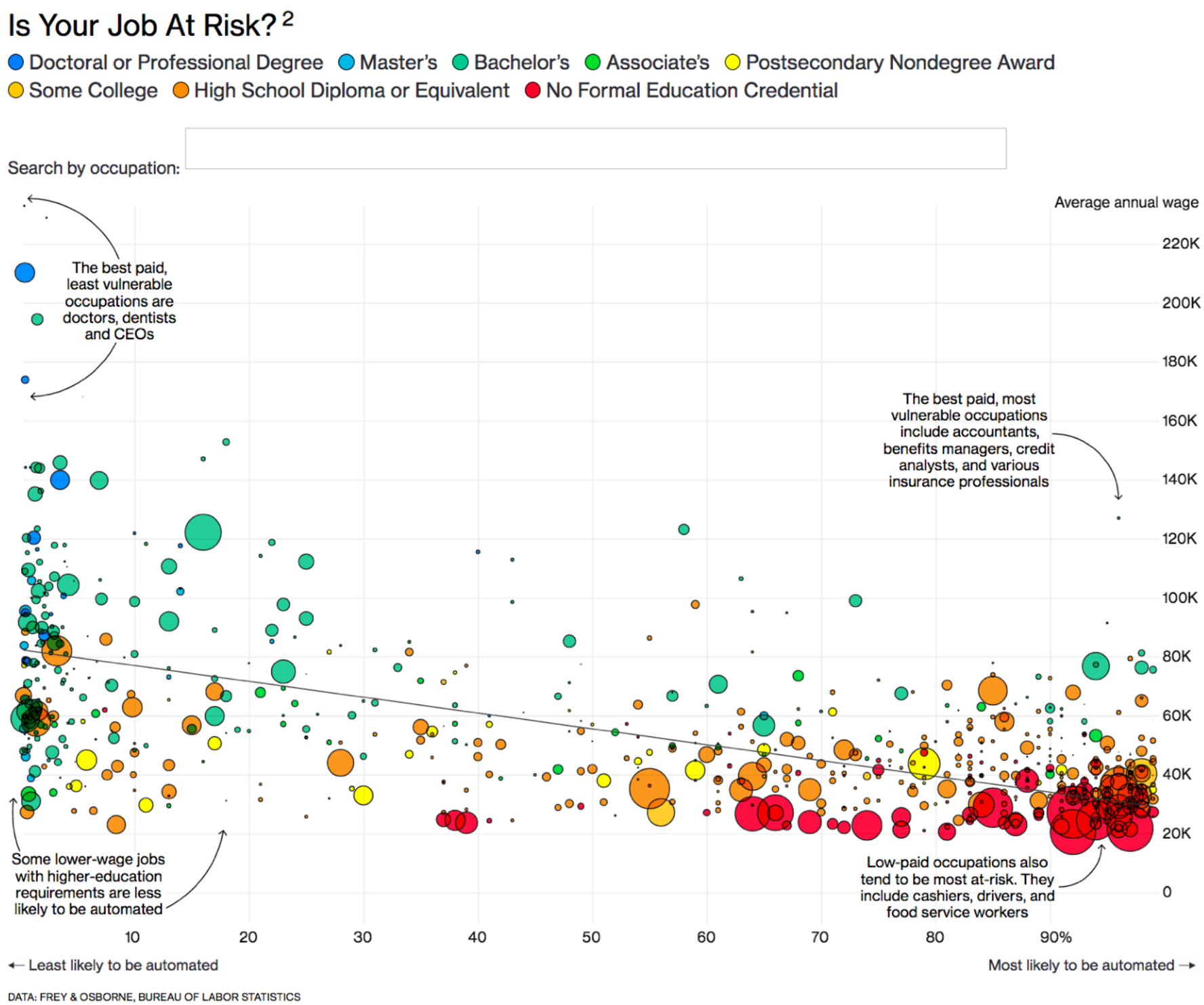
47% Of All Jobs Will Be Automated By 2034, And 'No Government Is Prepared' Says Economist

17/01/2014 14:43 | **Updated** 25 January 2014

2.4k

Michael Rundle
HuffPost UK Technology Editor

Artificial Intelligence



Laurent Alexandre : « L'Europe a complètement perdu la bataille de l'IA »

DAVID BARROUX | BENOIT GEORGES | 17/11/2017 |

Laurent Alexandre, chirurgien urologue, cofondateur du site web Doctissimo.fr, auteur du livre \'La defaite du cancer\' (éditions JC Lattes). - *RGA/REA*

+ VIEDO. Passionné, sûr de lui et volontiers provocateur, Laurent Alexandre occupe une place à part dans le petit monde des futuroloques français. Il fut l'un des premiers à alerter sur la puissance

Big Data

- GAFAM: Google, Apple, Facebook, Amazon, Microsoft
- Personal Information
  - Google, Facebook, Twitter, Linkedin,..
  - All personal communications in Europe are managed by non-European companies

Pour Cédric Villani, "la France n'a pas perdu la guerre de l'IA"

Par **Erick Haehnsen** | 07/12/2017, 8:33 | 1087 mots

Cédric Villani, Mathématicien et Député LRM de l'Essonne (Crédits : DR)

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.

AI Challenges

- La vision de la France doit donc consister à développer simultanément une IA **plus verte** et une IA au service de la transition écologique.

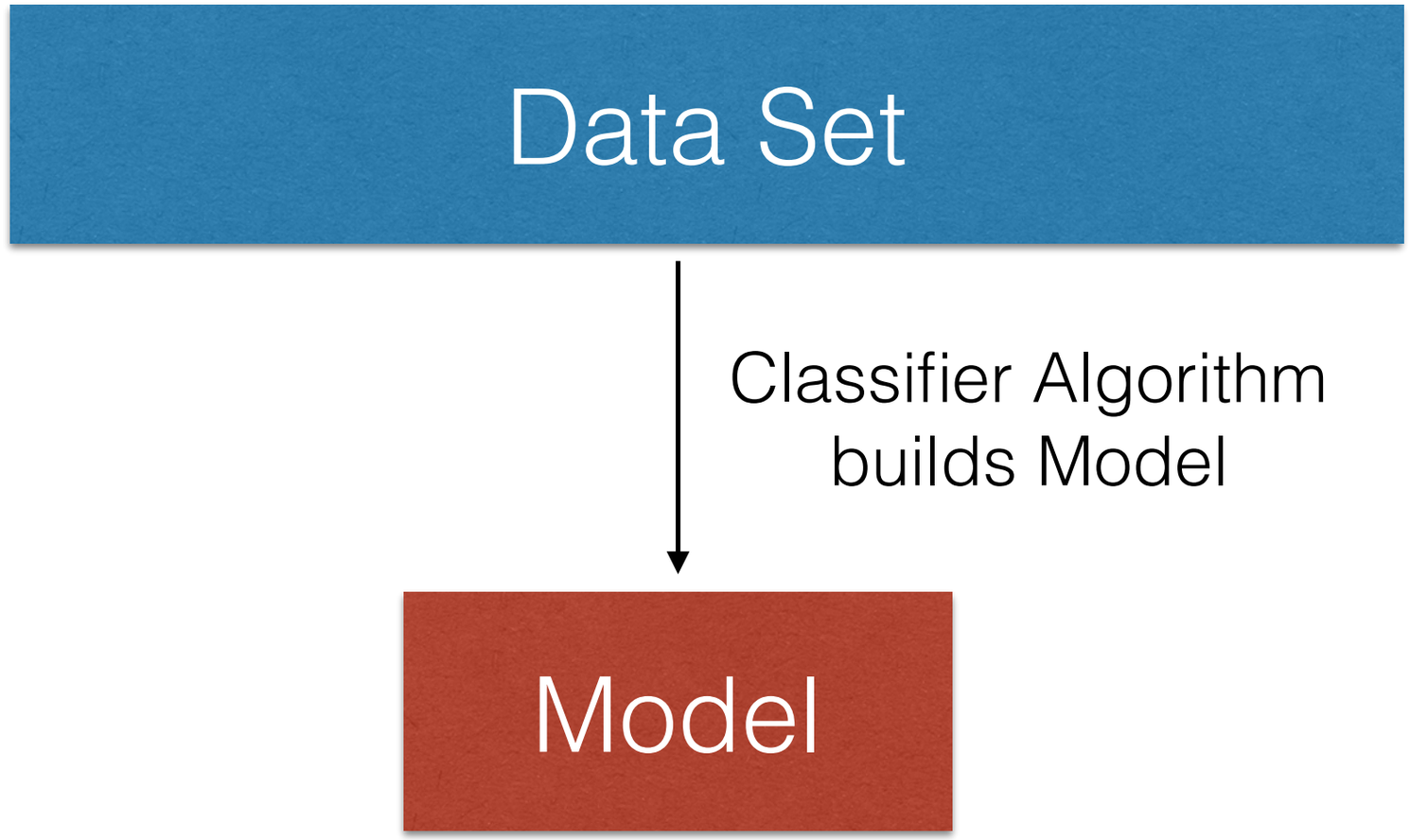
MACHINE LEARNING FOR DATA STREAMS

Institut Mines-Télécom

Machine Learning

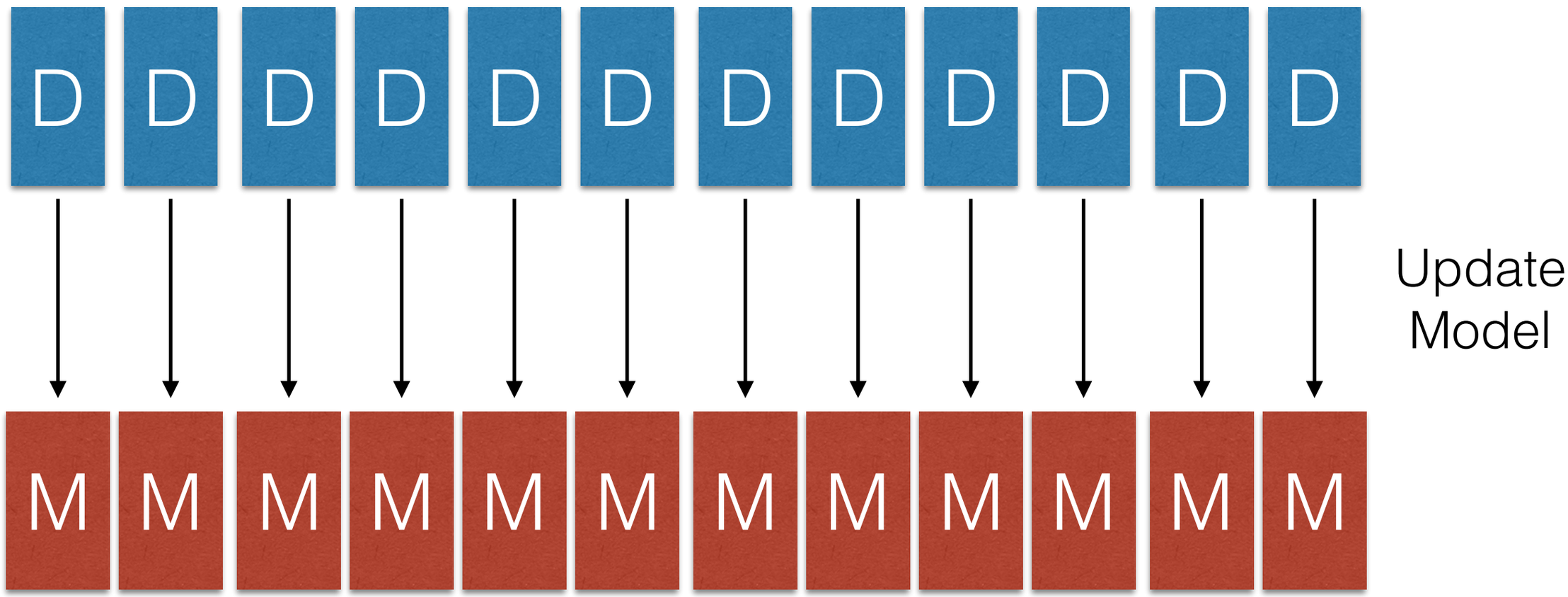
- **Machine learning** is a type of artificial intelligence (**AI**) that provides computers with the ability to learn without being explicitly programmed.

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



Analytic Standard Approach

Finite training sets
Static models

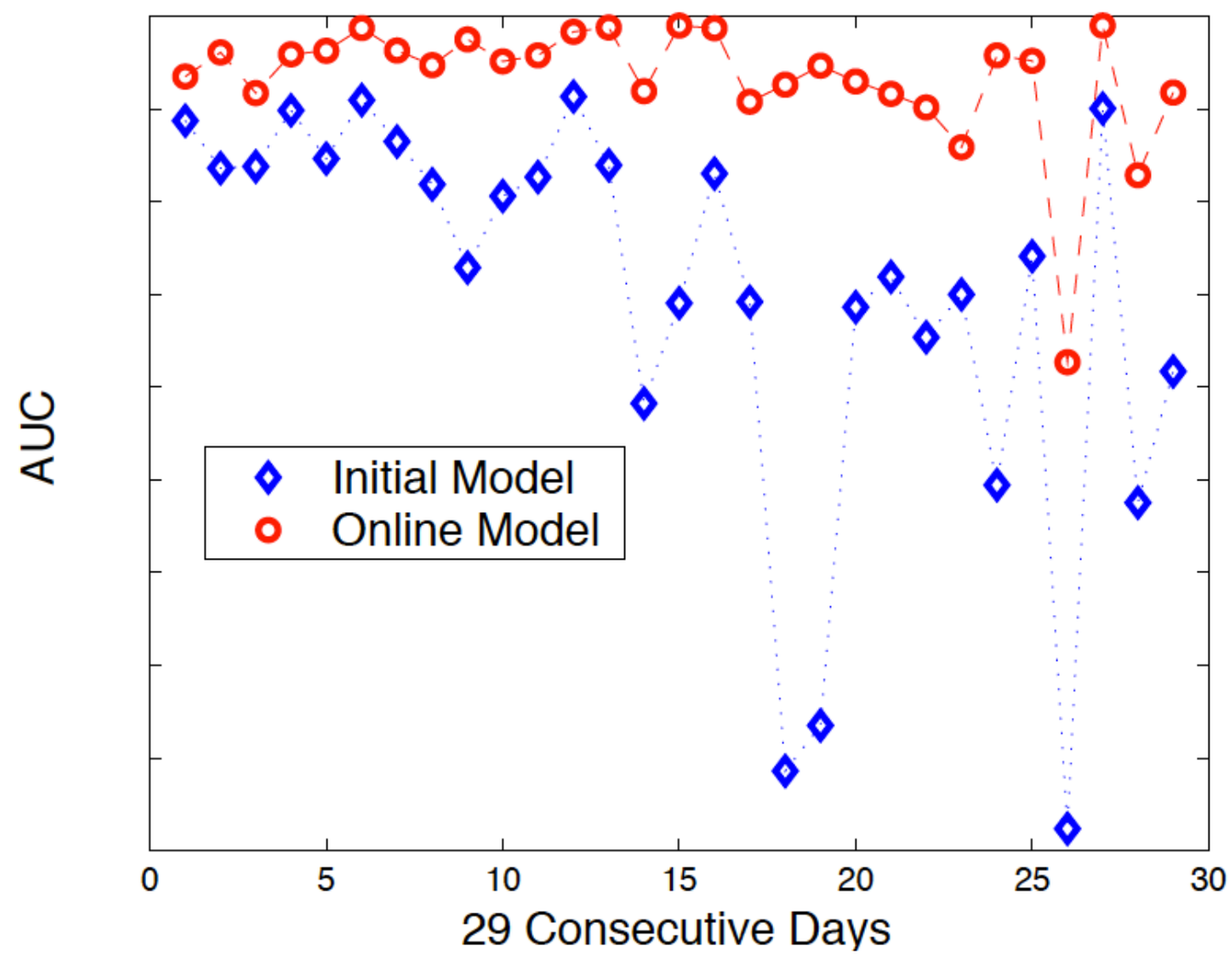


Data Stream Approach

Infinite training sets
Dynamic models

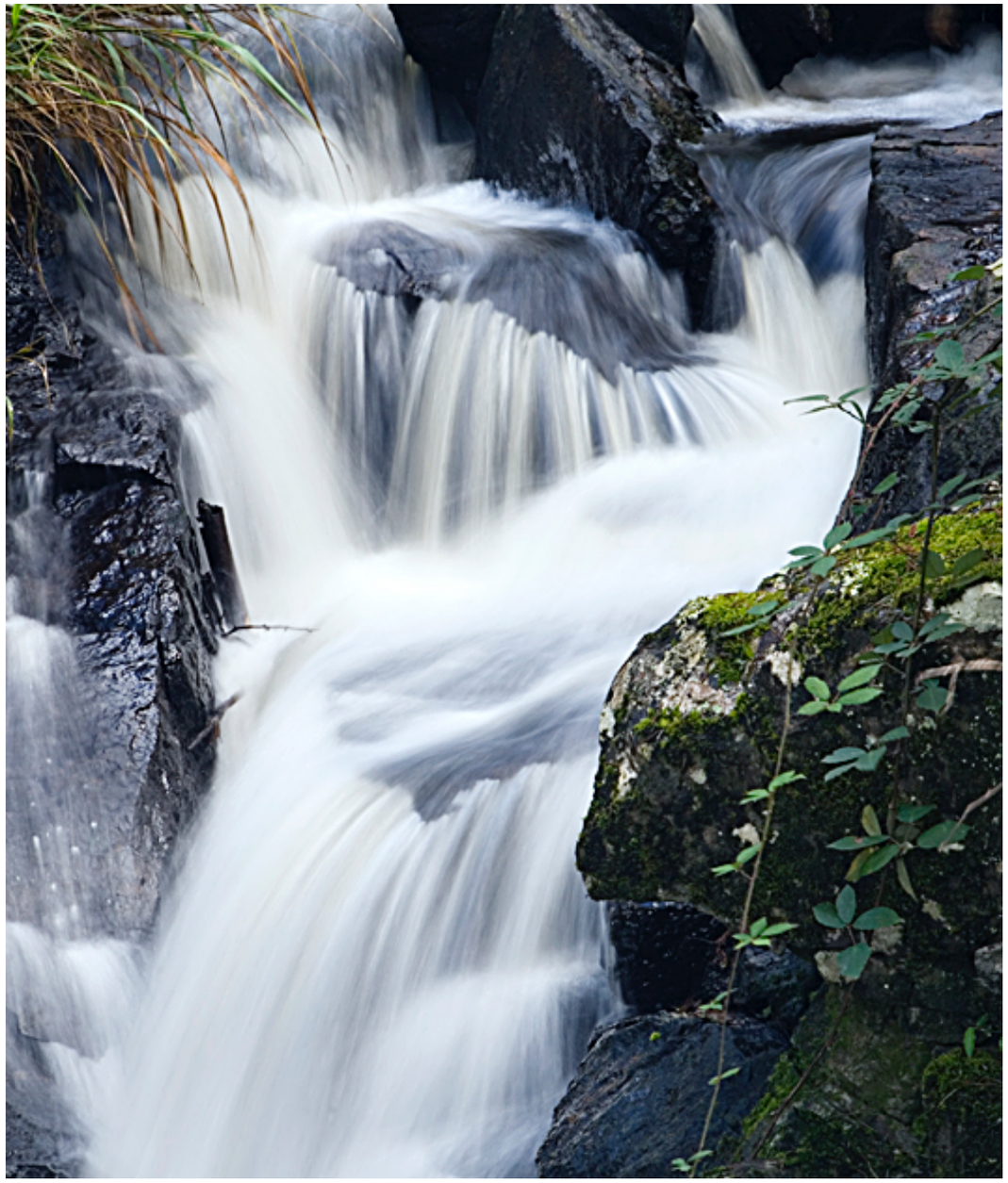
Pain Points

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



IoT Stream Mining

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



Approximation Algorithms

- General idea, good for streaming algorithms
- Small error ε with high probability $1-\delta$
  - True hypothesis H, and learned hypothesis $\hat{H}$
  - $\Pr[\ |H - \hat{H}| < \varepsilon|H|\ ] > 1-\delta$

A. Bifet, R. Gavaldà: “Adaptive Parameter-free Learning from Evolving Data Streams” IDA (2009)

Hoeffding Adaptive Tree

- Replace frequency counters by estimators
  - No need for window of instances
  - Sufficient statistics kept by estimators separately
- Parameter-free change detector + estimator with theoretical guarantees for subtree swap (ADWIN)
  - Keeps sliding window consistent with “no-change hypothesis”

A. Bifet, R. Gavaldà: “Learning from Time-Changing Data with Adaptive Windowing”. SDM ‘07

ADWIN

ADWIN

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

Problem

Given an input sequence $x_1, x_2, \ldots, x_t, \ldots$ we want to output

- a prediction $\widehat{x}_{t+1}$ minimizing prediction error:

$$|\widehat{x}_{t+1} - x_{t+1}|$$

- an alert if change is detected

ADWIN

Optimal Change Detector and Predictor

- High accuracy
- Fast detection of change
- Low false positives and false negatives ratios
- Low computational cost: minimum space and time needed

ADWIN

- Theoretical guarantees
- No parameters needed

ADWIN

Theorem

At every time step we have:

1. (False positive rate bound). *If μ_t remains constant within W, the probability that* ADWIN *shrinks the window at this step is at most δ.*
2. (False negative rate bound). *Suppose that for* some *partition of W in two parts $W_0 W_1$ (where W_1 contains the most recent items) we have $|\mu_{W_0} - \mu_{W_1}| > 2\epsilon_c$. Then with probability $1-\delta$* ADWIN *shrinks W to W_1, or shorter.*

ADWIN tunes itself to the data stream at hand, with no need for the user to hardwire or precompute parameters.

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)

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

OPEN SOURCE TOOLS

Institut Mines-Télécom

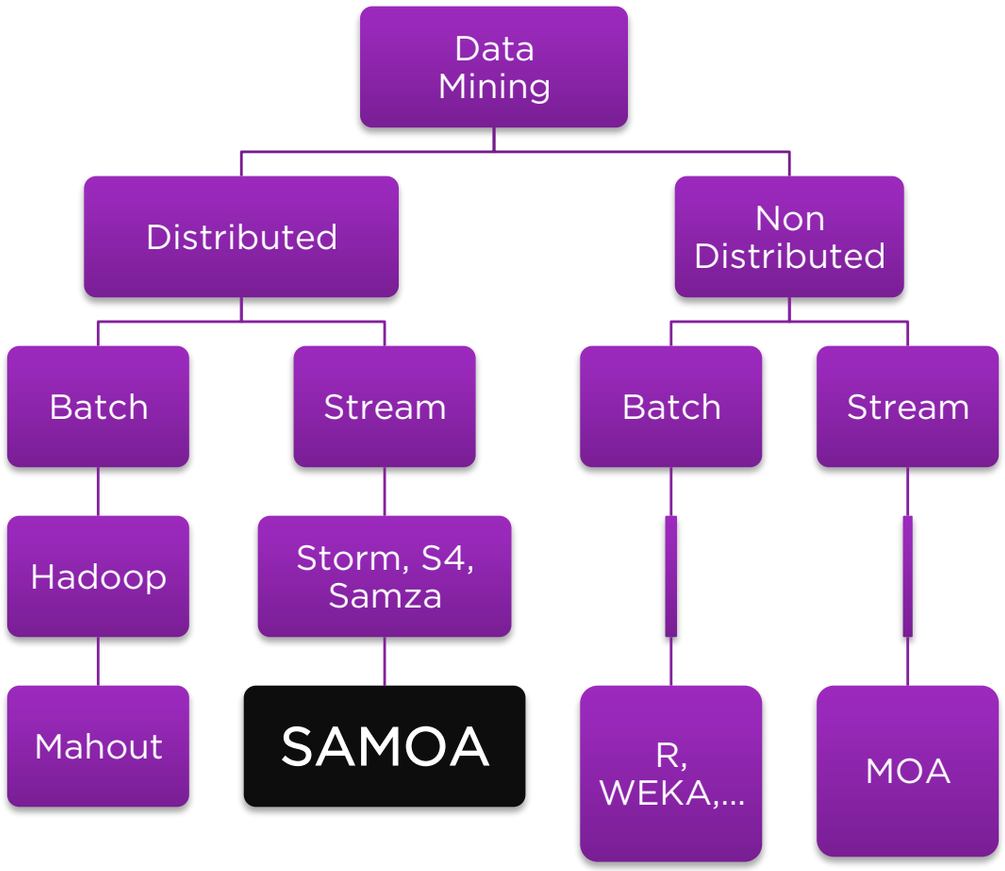
MOA

- {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

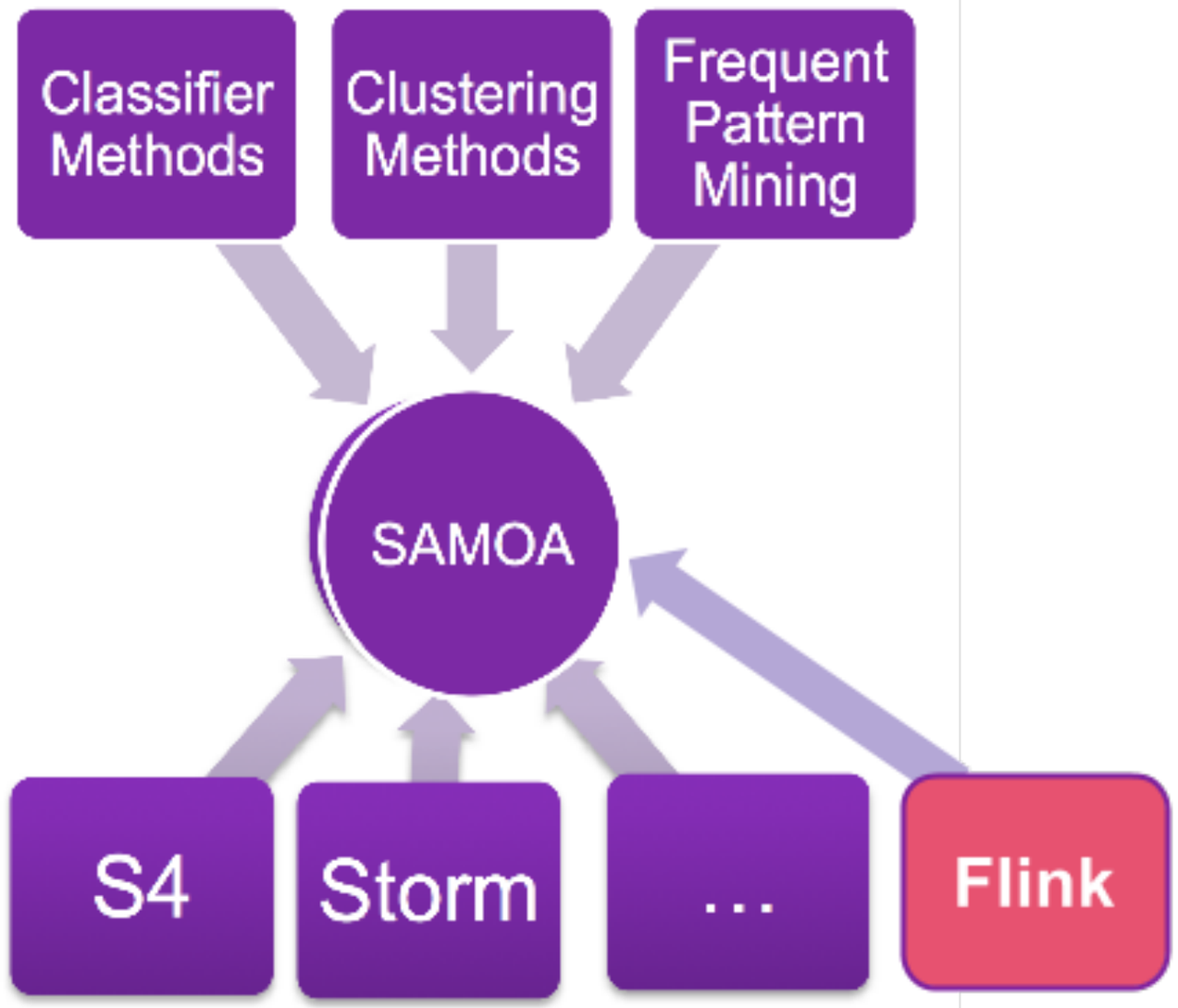
http://samoa-project.net

APACHE SAMOA

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

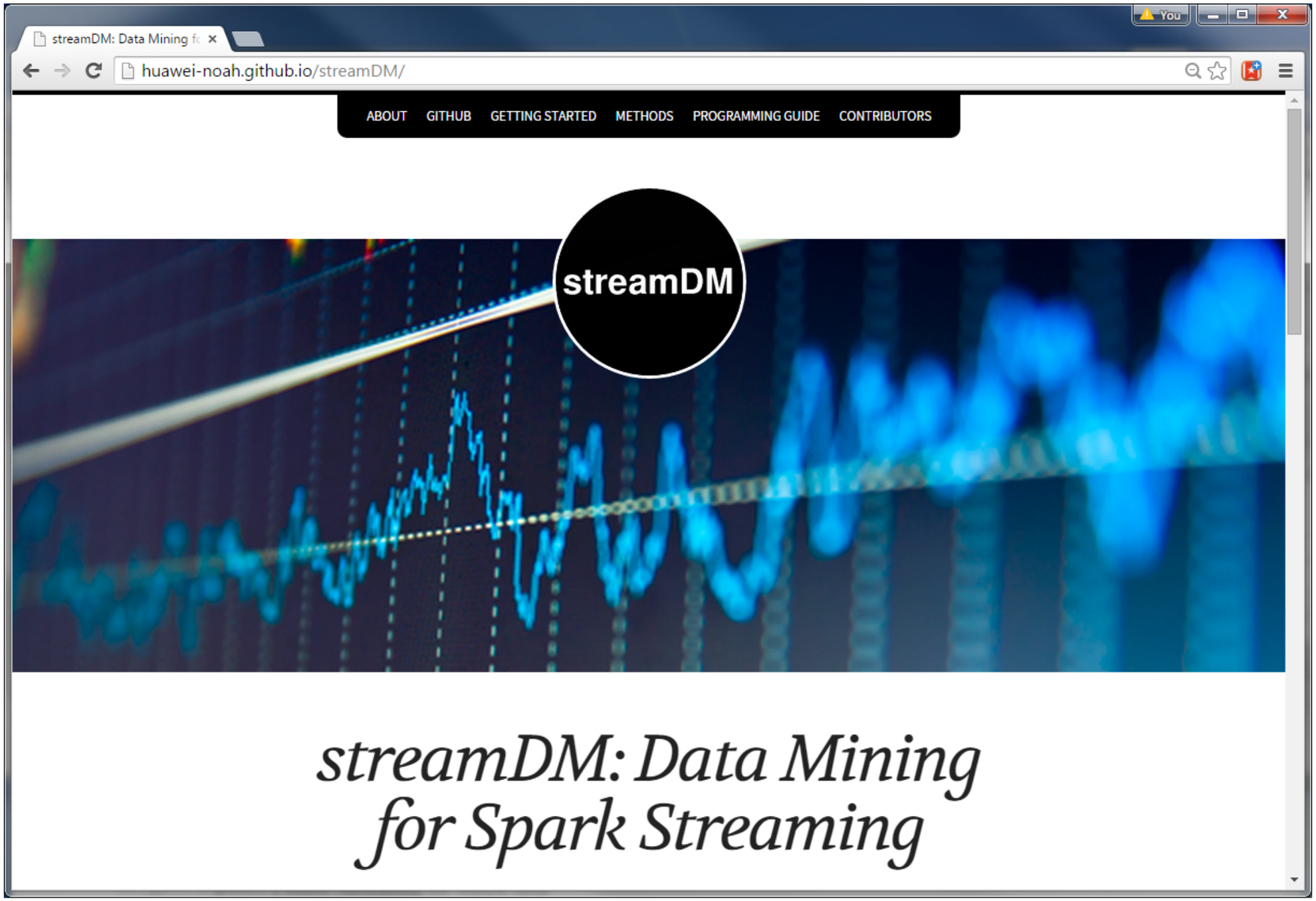


SAMOA ARCHITECTURE



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

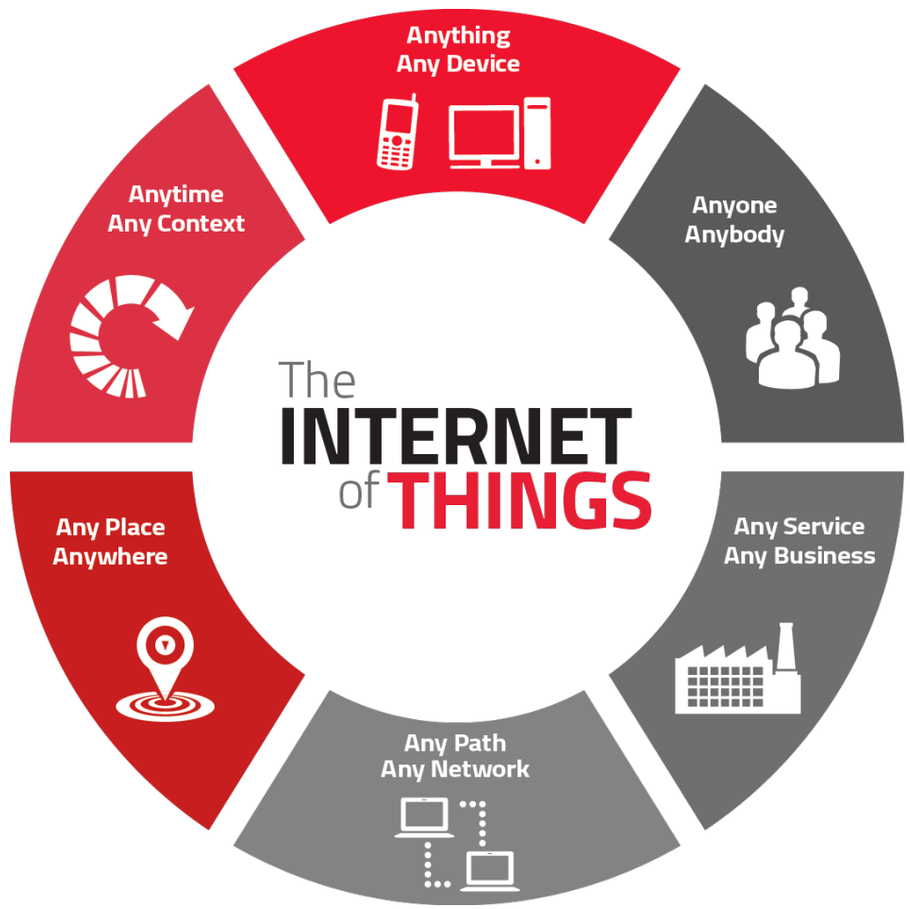
Mainly supported by: TELECOM ParisTech, ÉCOLE POLYTECHNIQUE UNIVERSITÉ PARIS-SACLAY

GITHUB

SUMMARY

Institut Mines-Télécom

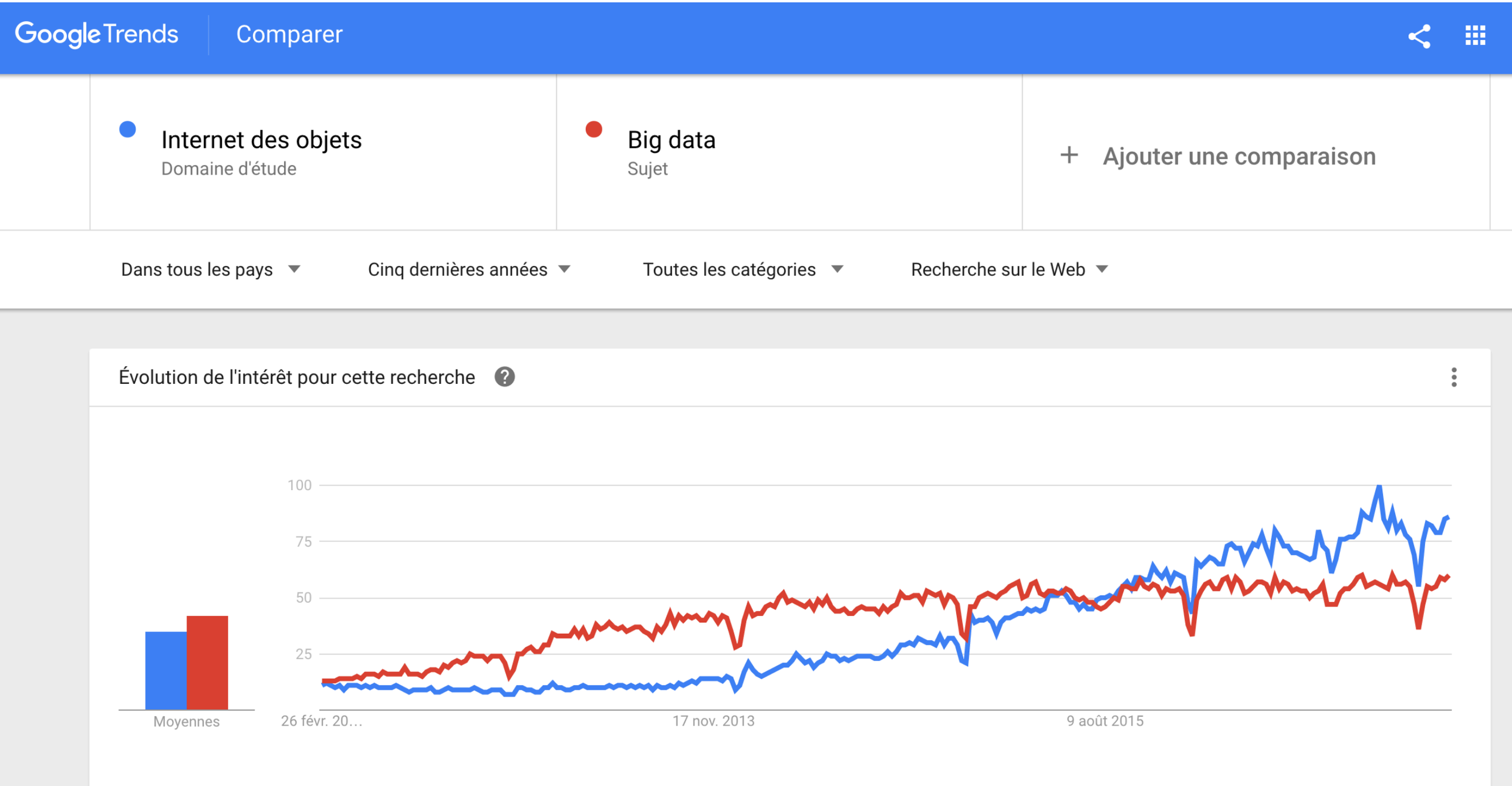
INTERNET OF THINGS



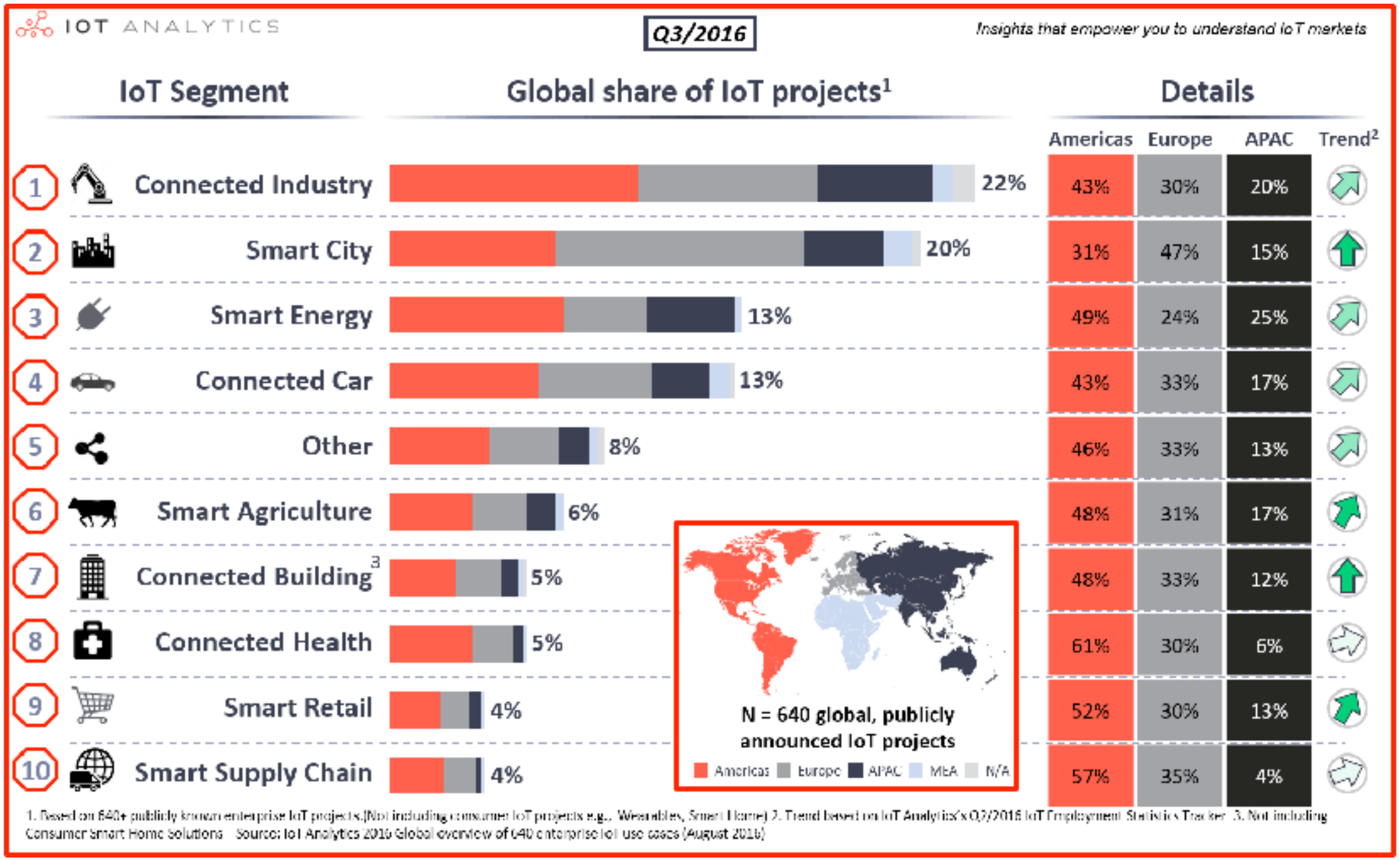
IoT: sensors and actuators connected by networks to computing systems.

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

IoT versus Big Data

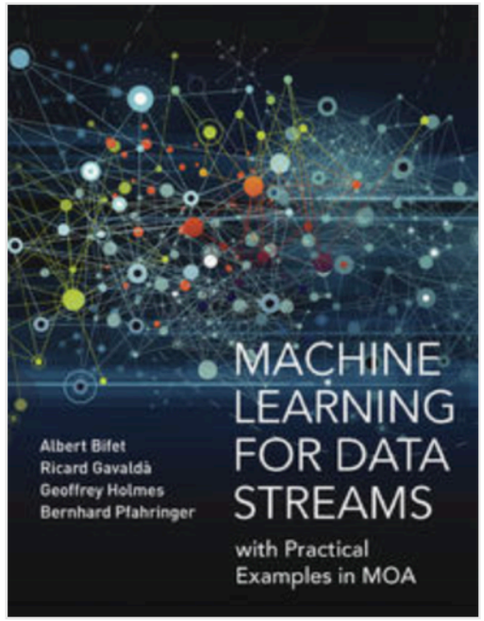


Applications IoT Analytics



IOT AND INDUSTRY 4.0

- Interoperability: IoT
- Information transparency: virtual copy of the physical world
- Technical assistance: support human decisions
- Decentralized decisions: make decisions on their own



Hardcover | **$55.00 Short** | **£45.95** | 288 pp. | 7 x 9 in | 21 color illus., 29 b&w illus. | February 2018 | ISBN: 9780262037792

Machine Learning for Data Streams

with Practical Examples in MOA

By Albert Bifet, Ricard Gavaldà, Geoff Holmes and Bernhard Pfahringer

Overview

Today many information sources—including sensor networks, financial markets, social networks, and healthcare monitoring—are so-called data streams, arriving sequentially and at high speed. Analysis must take place in real time, with partial data and without the capacity to store the entire data set. This book presents algorithms and techniques used in data stream mining and real-time analytics. Taking a hands-on approach, the book demonstrates the techniques using MOA (Massive Online Analysis), a popular, freely available open-source software framework, allowing readers to try out the techniques after reading the explanations.

The book first offers a brief introduction to the topic, covering big data mining, basic methodologies for mining data streams, and a simple example of MOA. More detailed discussions follow, with chapters on sketching techniques, change, classification, ensemble methods, regression, clustering, and frequent pattern mining. Most of these chapters include exercises, an MOA-based lab session, or both. Finally, the book discusses the MOA software, covering the MOA graphical user interface, the command line, use of its API, and the development of new methods within MOA. The book will be an essential reference for readers who want to use data stream mining as a tool, researchers in innovation or data stream mining, and programmers who want to create new algorithms for MOA.

Data, Intelligence and Graphs (DIG)

Home News Research Seminar Publications People Contact

HOME

Data, Intelligence and Graphs (DIG) is a group of researchers at LTCI, Télécom ParisTech who study the fundamental issues raised in data and **knowledge management** systems, **graph mining** and **artificial intelligence**. Research interests cover theoretical foundations of data intelligence and graph systems, practical solutions and applications, as well as cognitive aspects. The group was formerly known as DBWeb Team.

DIG has strong academic and industrial collaborations:

Research at Google SAFRAN CISCO orange™ NOKIA

Data, Intelligence and Graphs (DIG)

Senior

- **Talel Abdessalem**, Professor
- **Antoine Amarilli**, Associate Professor
- **Albert Bifet**, Professor, Head
- **Thomas Bonald**, Professor
- **Laurent Decreusefond**, Professor
- **Jean-Louis Dessalles**, Associate Professor
- **Yanlei Diao**, Professor (co-affiliated with École polytechnique)
- **Pierre Senellart**, Professeur Invité
- **Mauro Sozio**, Associate Professor
- **Fabian M. Suchanek**, Professor

WIKIPÉDIA
L'encyclopédie libre

Accueil
Portails thématiques
Article au hasard
Contact

Contribuer
Débuter sur Wikipédia
Aide
Communauté
Modifications récentes
Faire un don

Outils
Pages liées
Suivi des pages liées
Importer un fichier
Pages spéciales
Lien permanent
Informations sur la page
Élément Wikidata
Citer cette page

Jean-Louis Dessalles

Cet article est une ébauche concernant les sciences cognitives.
Vous pouvez partager vos connaissances en l'améliorant (**comment ?**) selon les recommandations des projets correspondants.

Pour les articles homonymes, voir Dessalles.

Jean-Louis Dessalles, né le 1er janvier 1956 à Périgueux, est un informaticien français, chercheur en Intelligence Artificielle et en Sciences cognitives, professeur à Télécom ParisTech, connu notamment pour ses contributions à la Théorie de la simplicité et pour une théorie originale sur l'origine (politique) du langage.

Jean-Louis Dessalles	
Naissance	1er janvier 1956 Périgueux, Dordogne, France
Nationalité	française
Domaines	Sciences cognitives
Institutions	Télécom ParisTech
Diplôme	École polytechnique (France) Télécom ParisTech
Renommé pour	Théorie de la simplicité Intelligence artificielle Origine du langage

modifier

Sommaire [masquer]

Biographie [modifier | modifier le code]

YAGO

Cet article est une ébauche concernant le Web sémantique.
Vous pouvez partager vos connaissances en l'améliorant (**comment ?**) — pour plus d'indications, visitez le projet Web Sémantique.

YAGO (Yet Another Great Ontology) est une base de connaissance créée par l'institut Max-Planck d'informatique à Sarrebruck. Elle est constituée à partir d'informations extraites de Wikipédia et d'autres sources.

En 2012, YAGO2s, la deuxième version de YAGO, possède 10 millions d'entités avec plus de 120 millions d'informations à propos de ces entités[2]. Les connaissances de YAGO sont extraites de Wikipédia (catégories, redirections, infoboxs), de WordNet (synsets, hyponymie), et de GeoNames[3]. La fiabilité de YAGO a été manuellement évaluée sur un échantillon de faits et est supérieure à 95%[4]. Afin de l'intégrer au web des données, YAGO est liée aux ontologies DBpedia[5] et SUMO[6].

YAGO2s est fourni dans la syntaxe Turtle et au format tsv. Des sauvegardes totales ou dumps de la base de données sont disponibles, des sauvegardes partielles thématiques ou spécialisées sont également disponible. Des requêtes peuvent être lancées à partir de

YAGO	
Développeur	Institut Max-Planck d'informatique à Sarrebruck
Première version	2008
Dernière version	YAGO3 (24 mars 2015)
Écrit en	Perl et Prolog
Type	Web sémantique, Web des données
Licence	Creative Commons CC-BY 3.0[1]
Site web	www.mpi-inf.mpg.de [archive]

Thanks!

@abifet

Institut Mines-Télécom

Futur & Ruptures a 10 ANS ! 2008-2018

FONDATION Mines-Télécom La Fondation de l'IMT

INSTITUT CARNOT Télécom & Société numérique

ARTIFICIAL INTELLIGENCE AND DATA STREAM MINING

ALBERT BIFET

15 FEVRIER 2018

#Futur&Ruptures

IMT Atlantique Bretagne-Pays de la Loire École Mines-Télécom · IMT Lille Douai École Mines-Télécom IMT-Université de Lille · IMT Mines Albi-Carmaux École Mines-Télécom · IMT Mines Alès École Mines-Télécom · MINES Saint-Étienne Une école de l'IMT · TELECOM École de Management Une école de l'IMT · TELECOM ParisTech Une école de l'IMT · TELECOM SudParis Une école de l'IMT · EURECOM · MINES ParisTech · MINES nancy · ARMINES