Introduction to IoT Data Stream Mining

Albert Bifet (@abifet)

Paris, 22 November 2017
albert.bifet@telecom-paristech.fr

Who are We

- Jesse Read
  - Associate Professor at École Polytechnique
  - MultiLabel Learning, Data stream mining and Deep Learning
    - MEKA: Multilabel Learning
    - MOA: Massive Online Analytics
- Albert Bifet
  - Associate Professor at Télécom ParisTech
  - Data stream mining algorithms and systems
    - MOA: Massive Online Analytics
    - Apache SAMOA: Scalable Advanced Massive Online Analytics

IoT Data Stream Mining

Outline

1. Introduction
2. Stream Algorithmics
3. Classification in Multi-output Data Streams
4. Concept Drift
5. Multi-output Learning
6. Ensemble Methods
7. Regression
8. Clustering
9. Frequent Pattern Mining

IoT Data Stream Mining

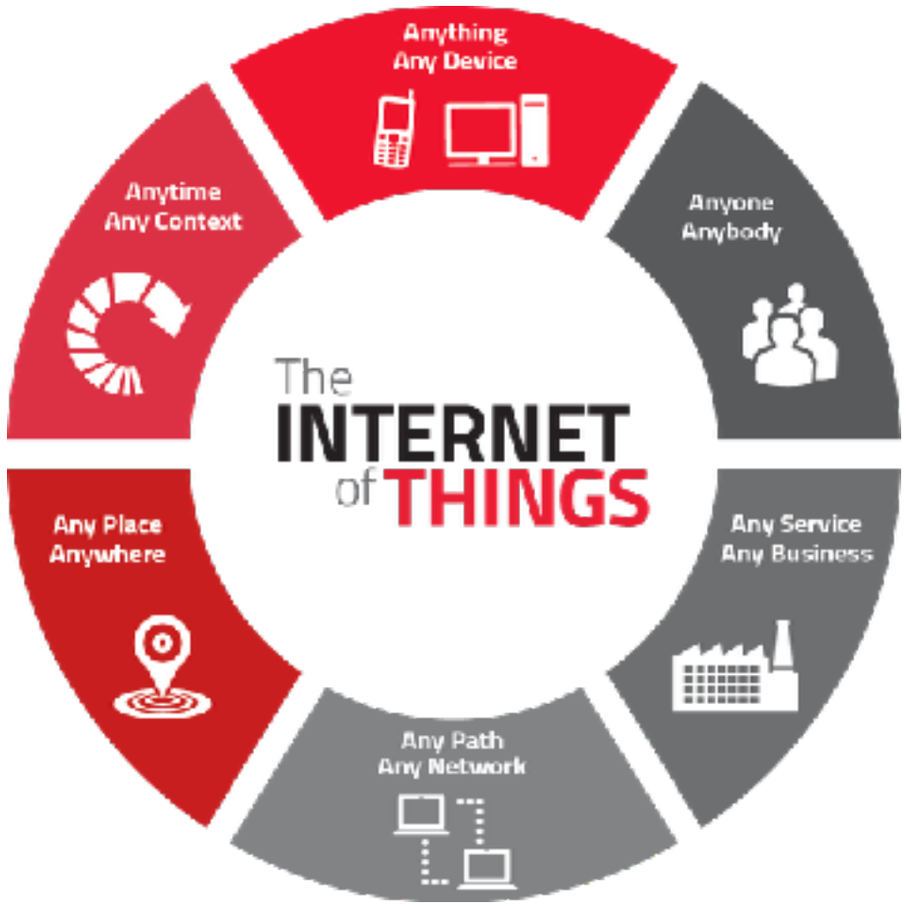
Assessment
10% Lab Assignments
30% Project
60% Test

Classes
22/11, 29/11, 13/12, 20/12 Wednesdays at 9:00

Session Labs: 6/12 and 10/01

Important Dates
Project Presentation: January 17
Test: January 31

INTERNET OF THINGS

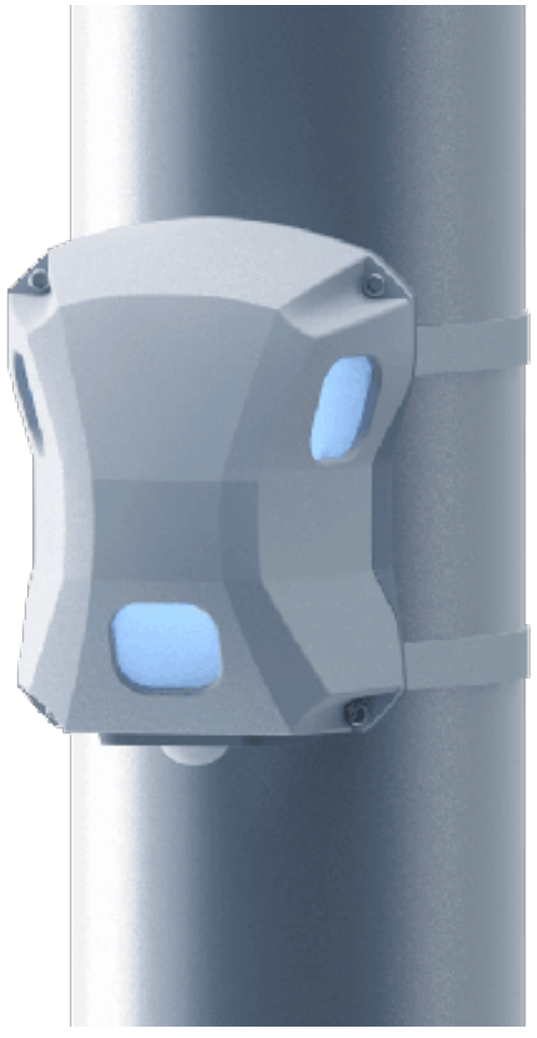


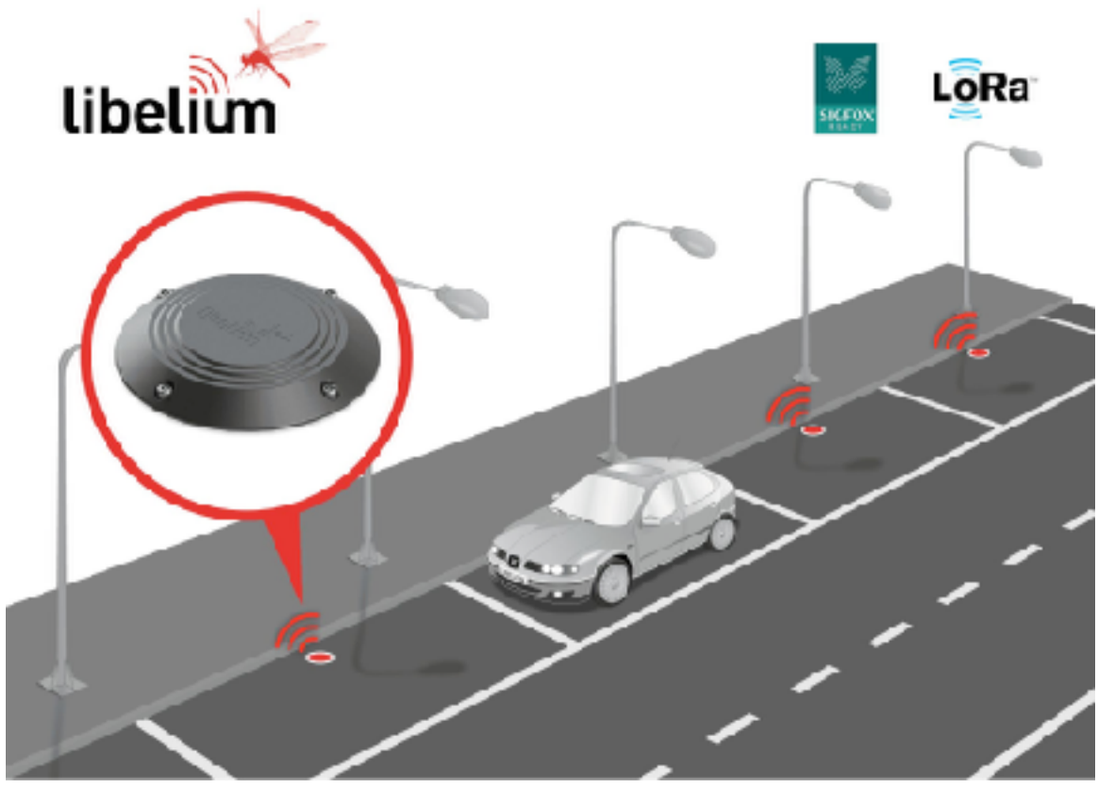
IoT: sensors and actuators connected by networks to computing systems.

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

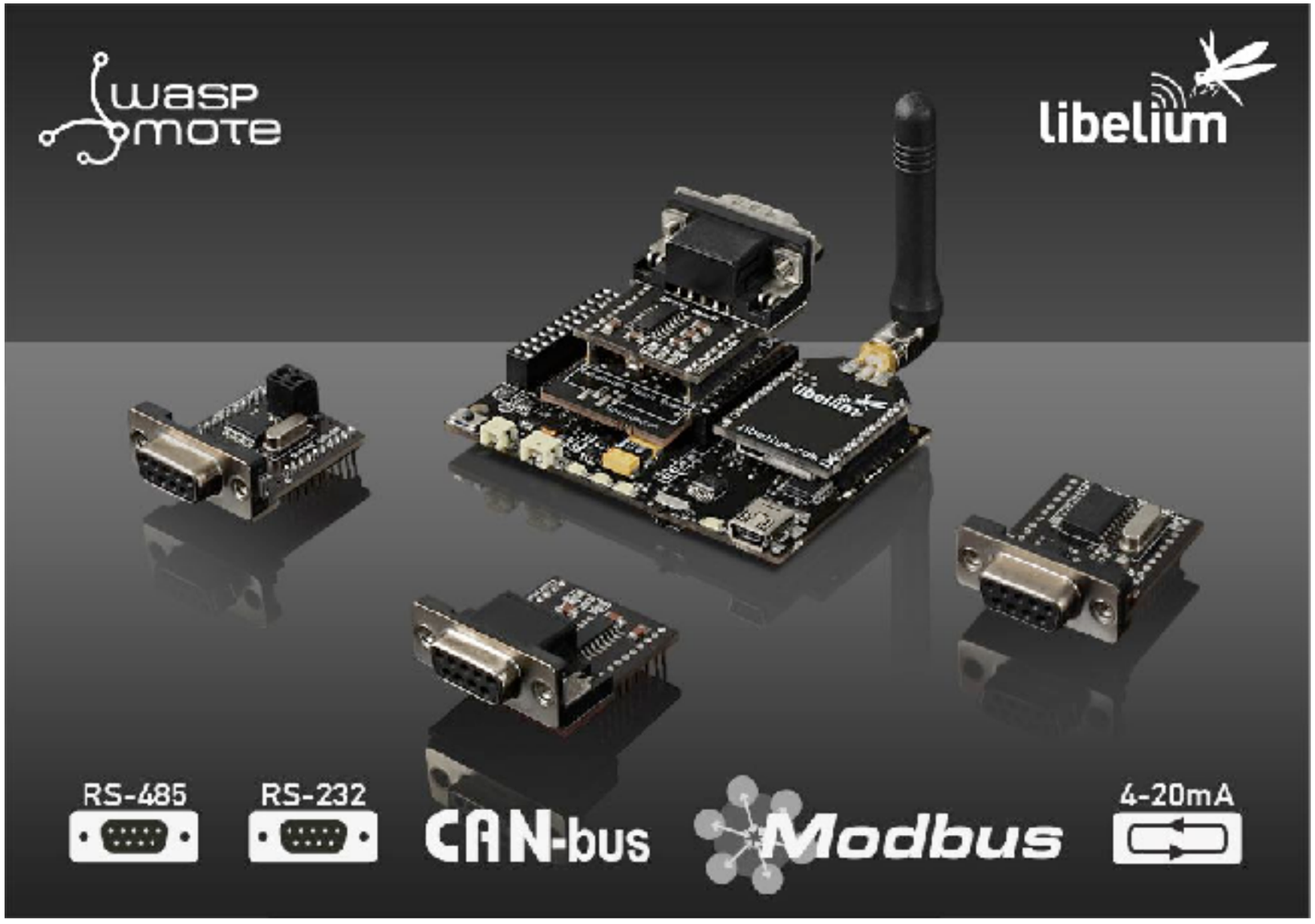
IoT Applications For Energy Management

IoT Applications For Connected/Smart Home



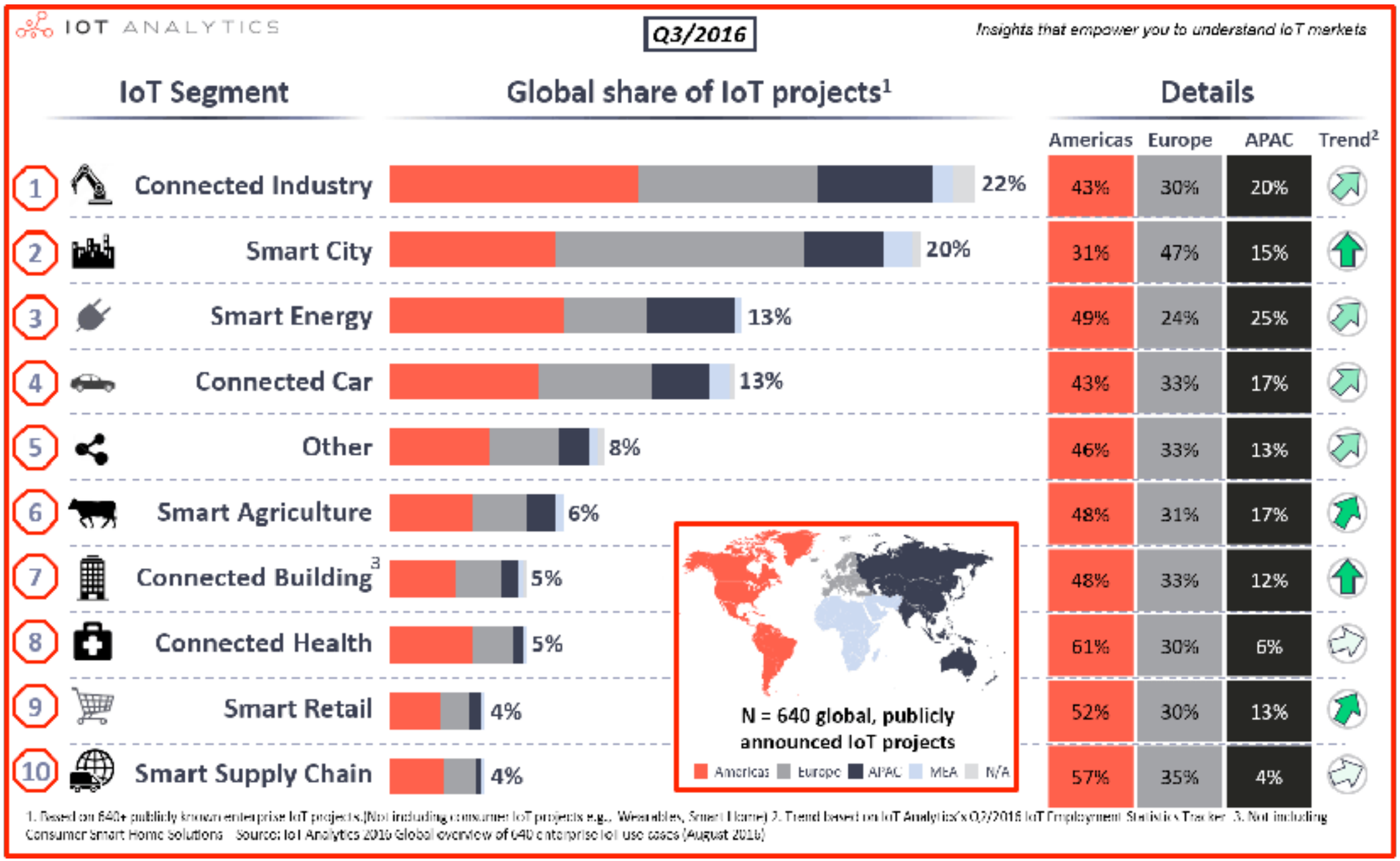


IoT Applications For Smart Cities



IoT Applications
For Industrial Automation

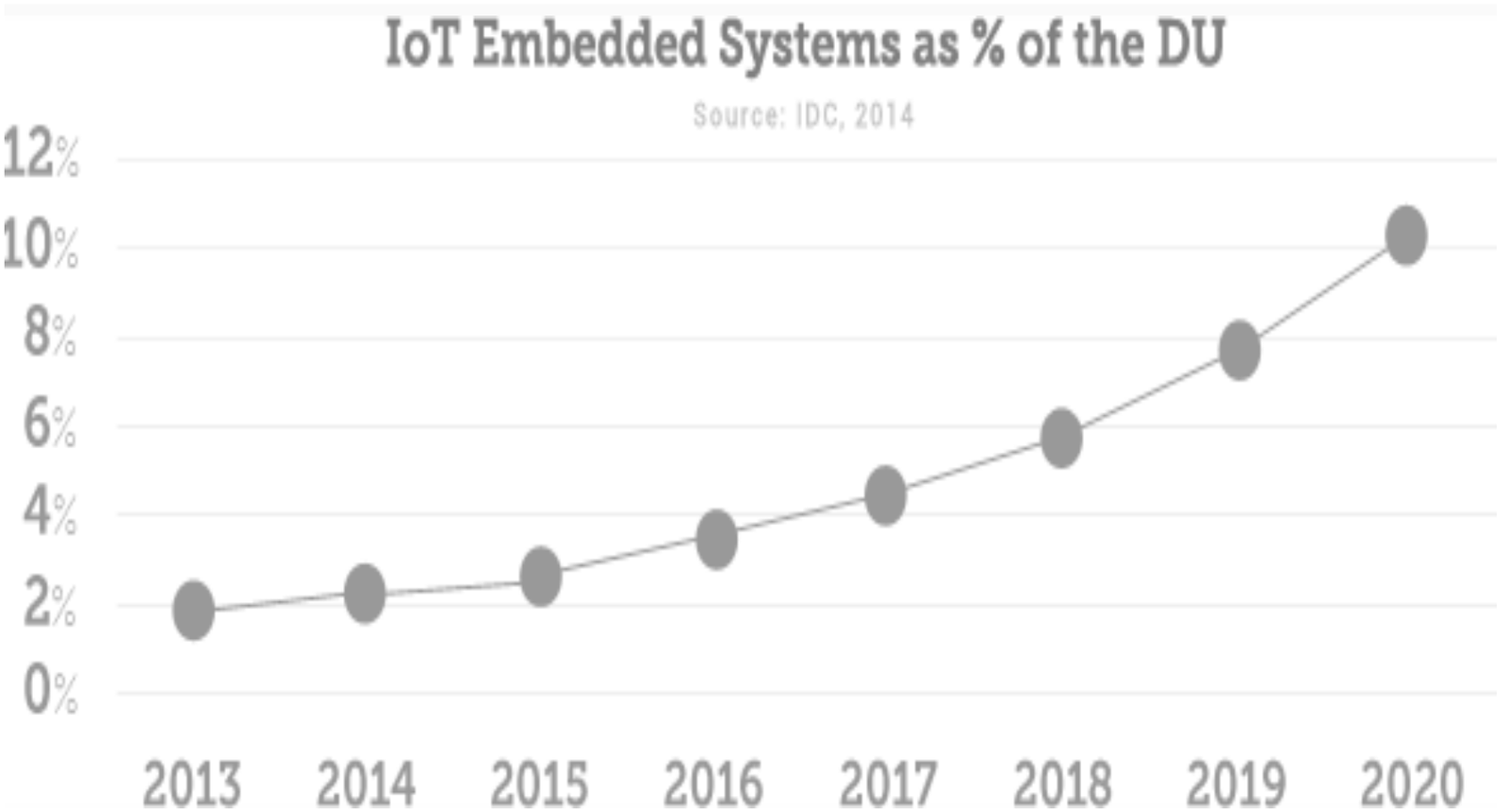
Applications IoT Analytics



IOT AND INDUSTRY 4.0

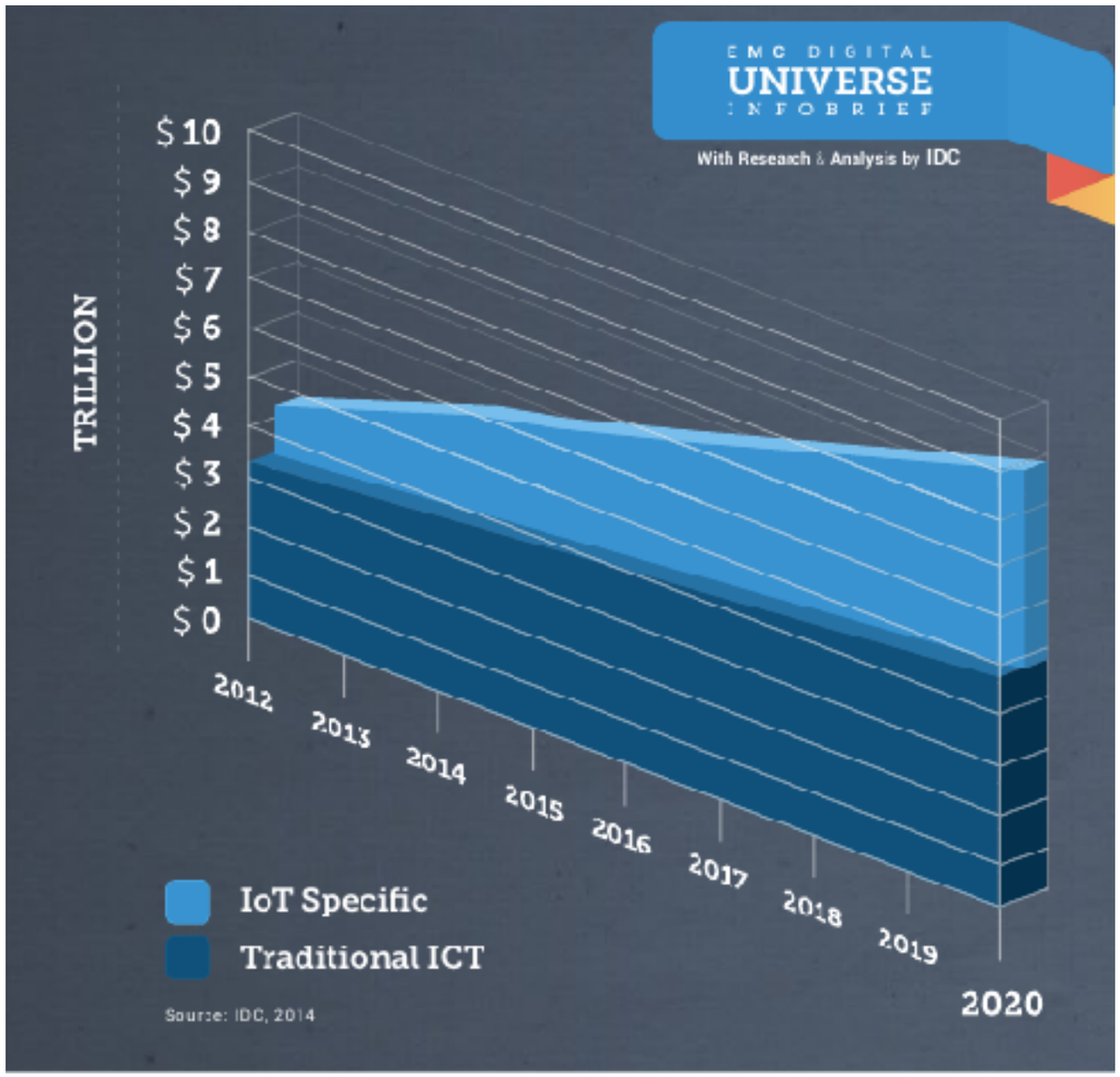
- Interoperability: IoT
- Information transparency: virtual copy of the physical world
- Technical assistance: support human decisions
- Decentralized decisions: make decisions on their own

INTERNET OF THINGS

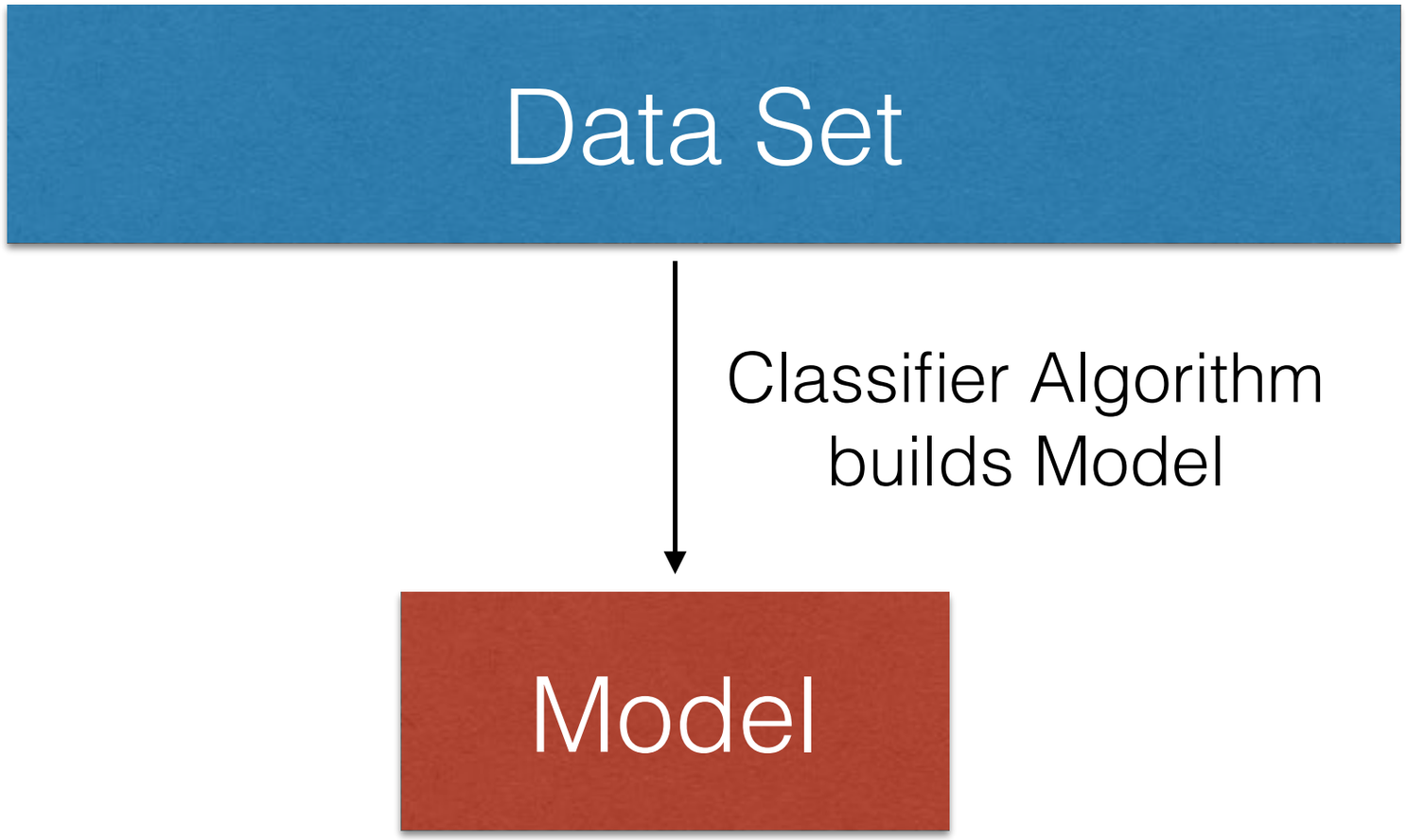


- EMC Digital Universe, 2014

INTERNET OF THINGS

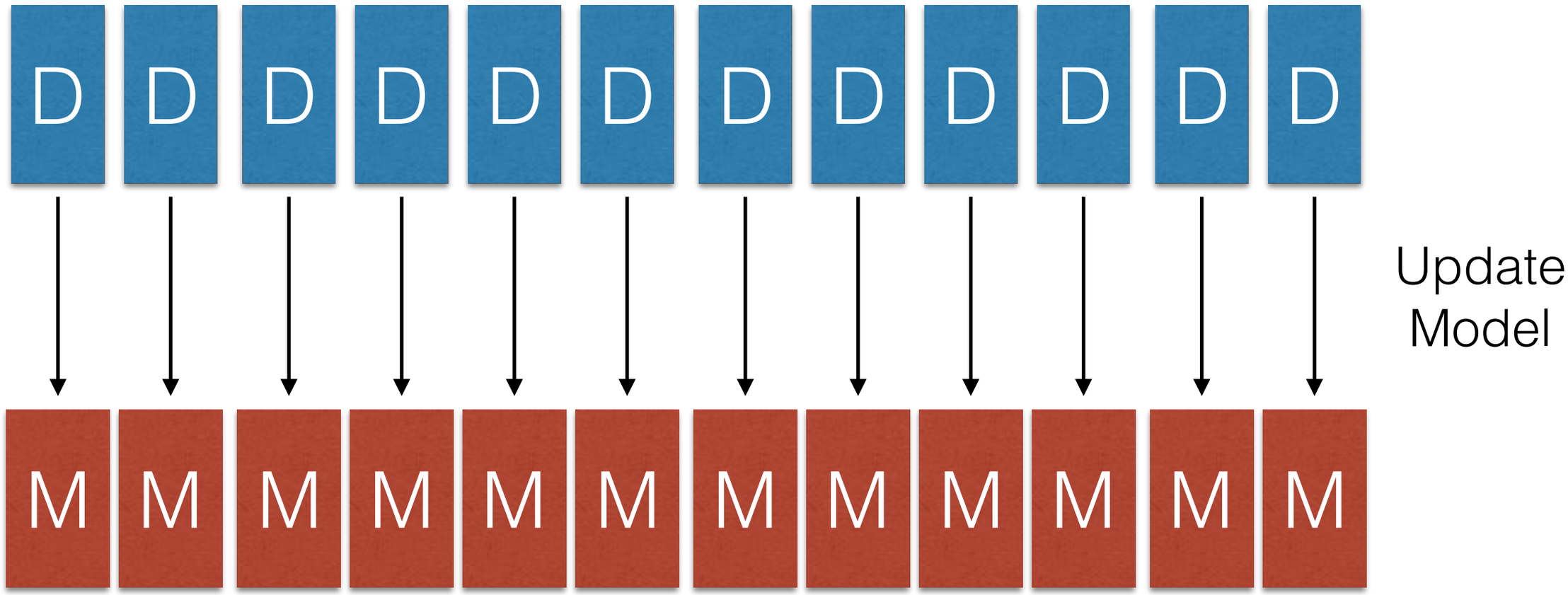


- EMC Digital Universe, 2014



Analytic Standard Approach

Finite training sets
Static models

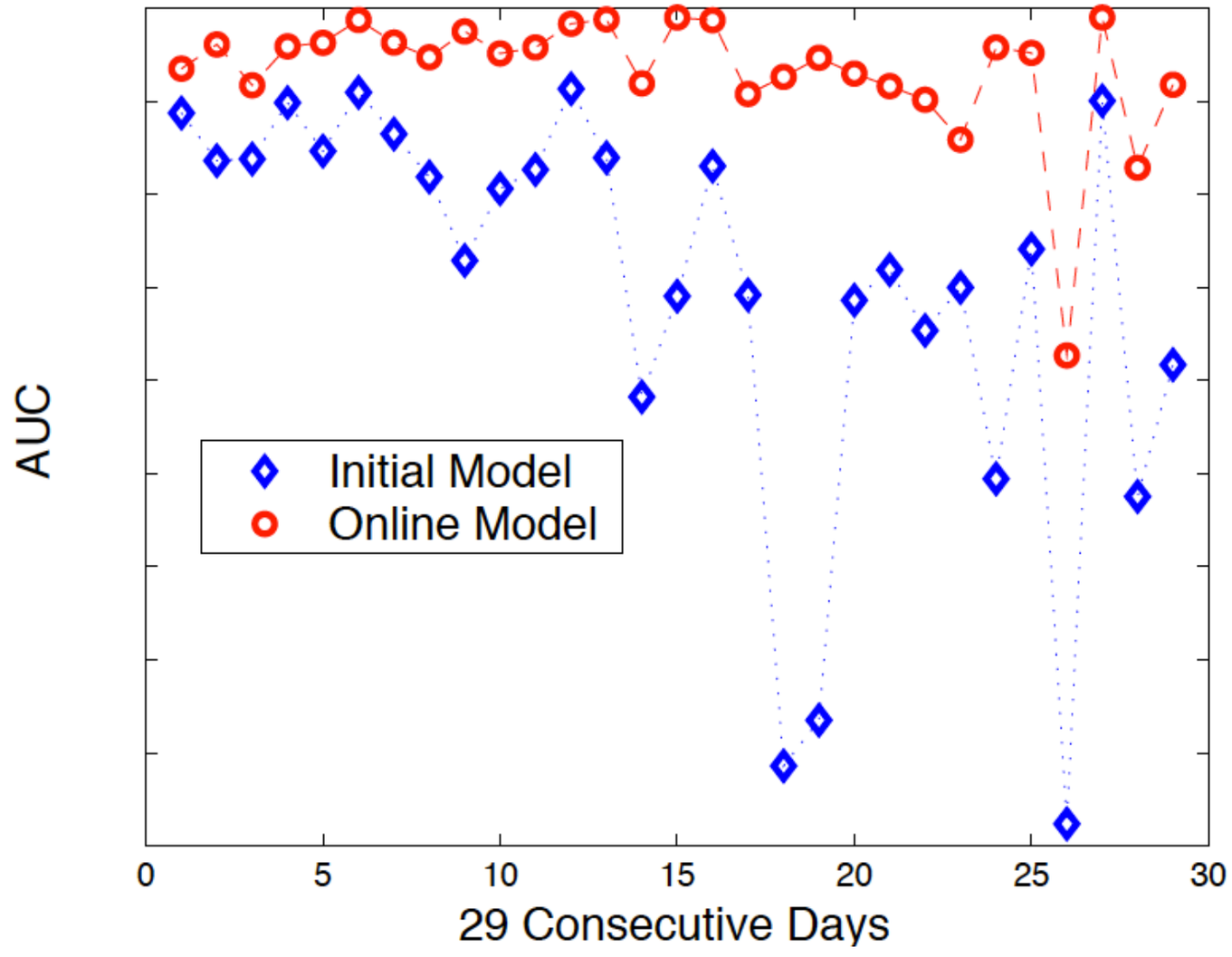


Data Stream Approach

Infinite training sets
Dynamic models

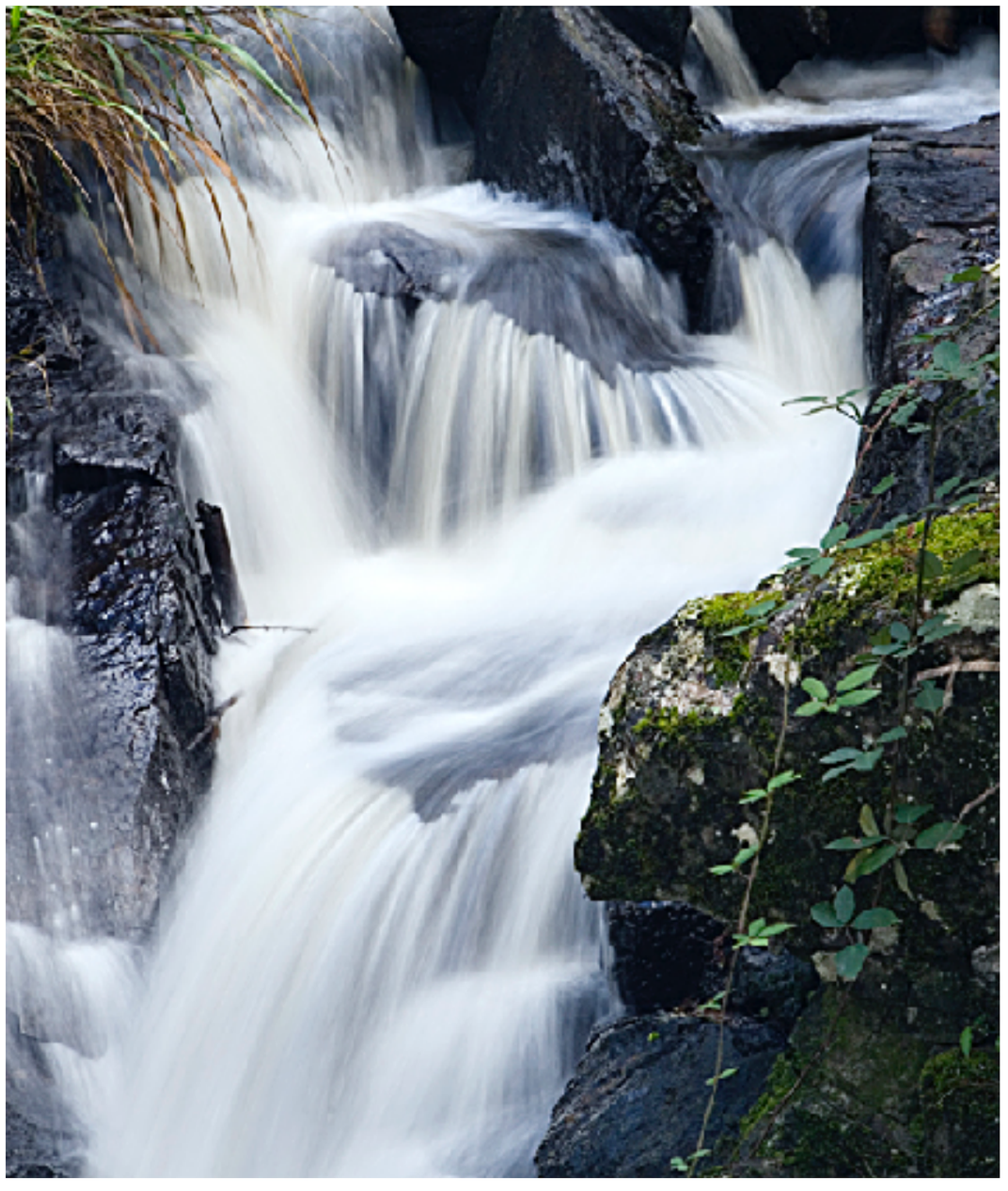
Pain Points

- 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

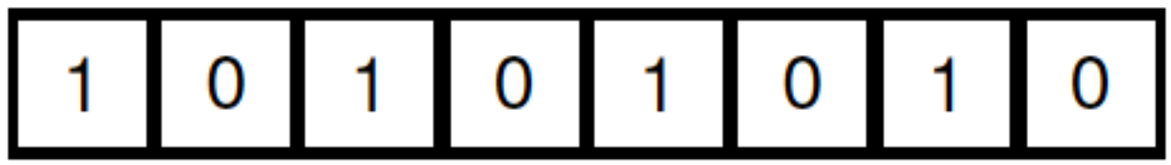


Approximation Algorithms

- General idea, good for streaming algorithms
- Small error ε with high probability 1-δ
  - 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?



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.

WHAT IS **MOA**?

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

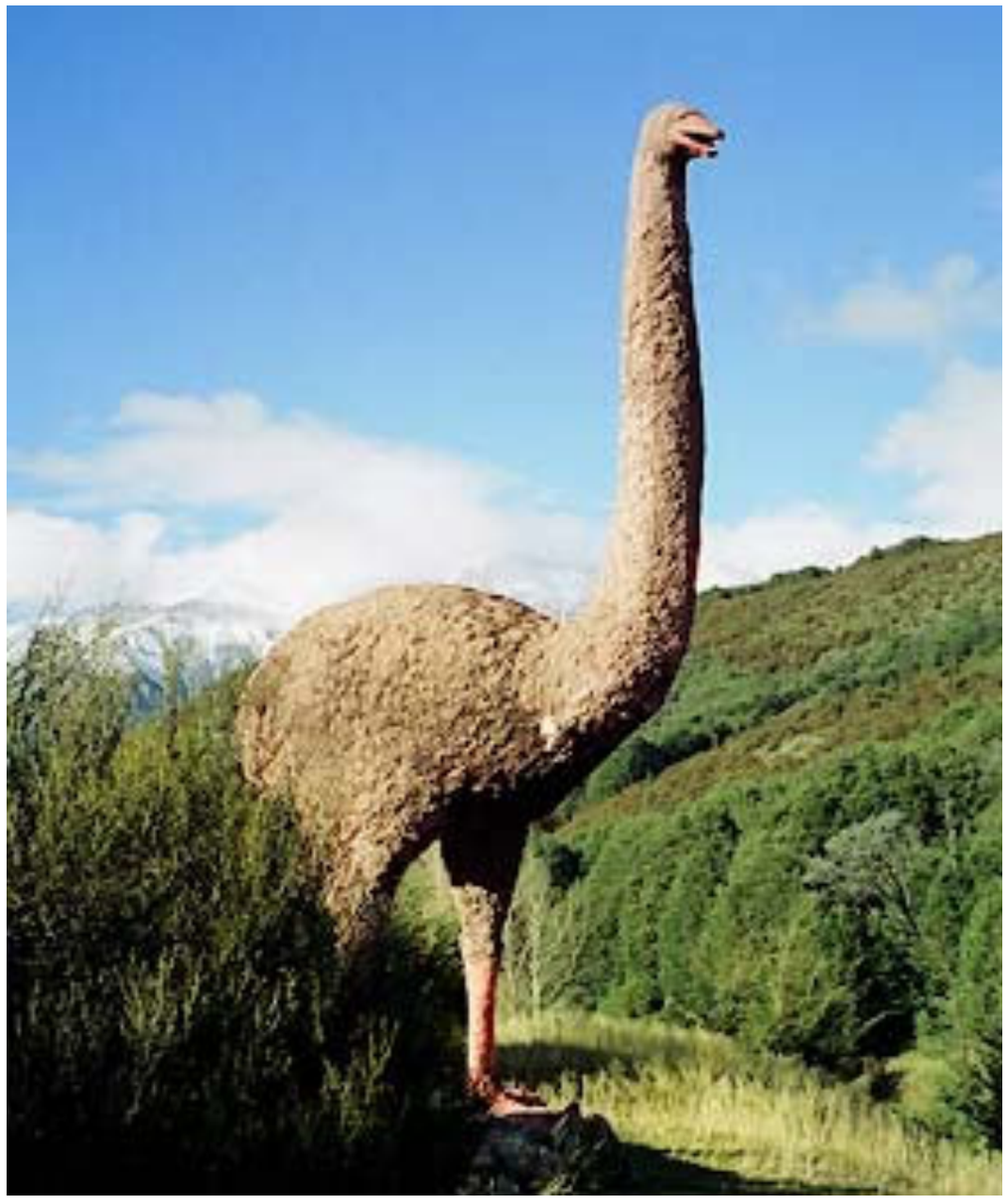
  - clustering, frequent pattern mining

- Easy to extend, design and run experiments

WEKA: the bird

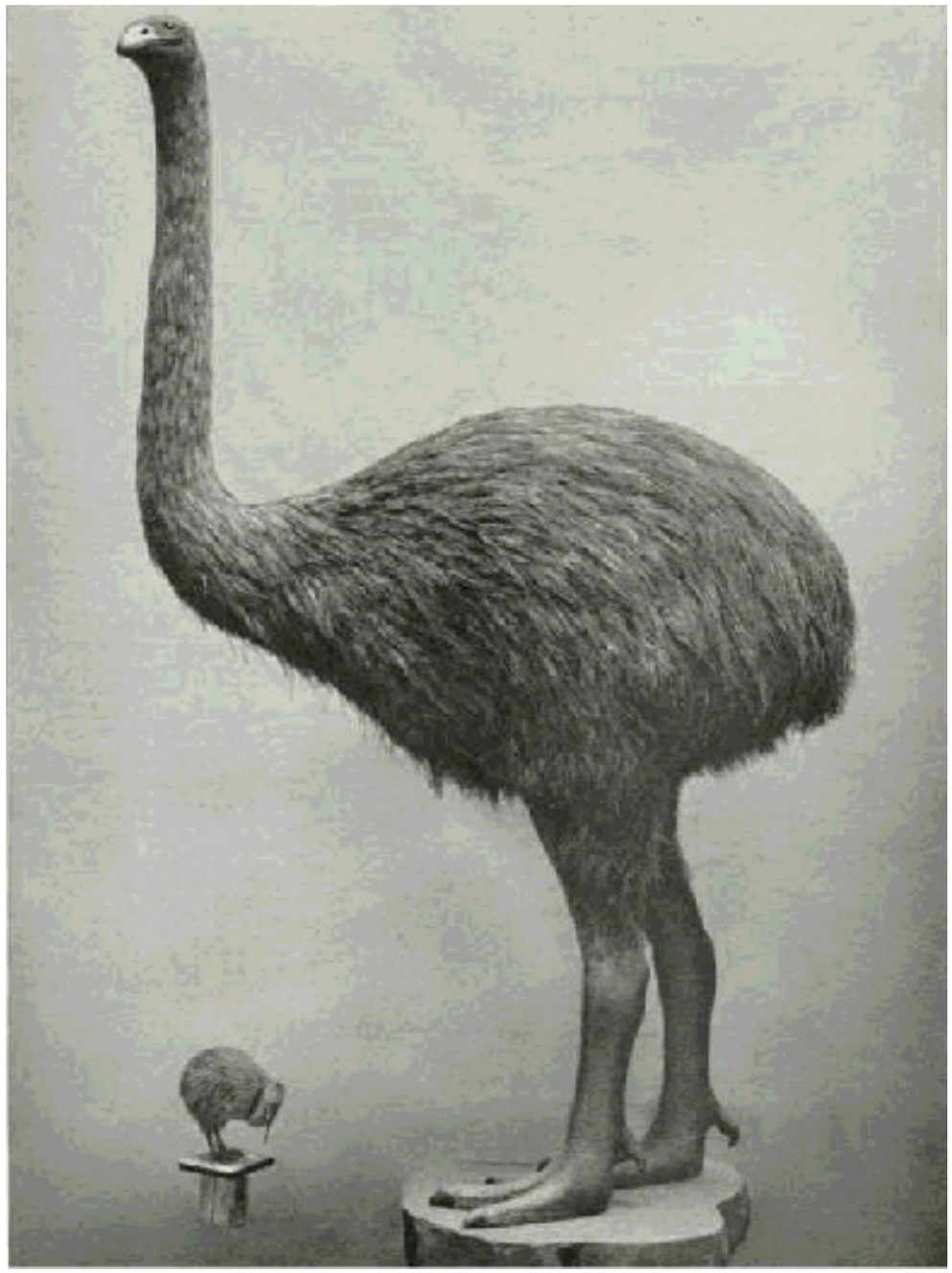
MOA: the bird

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.

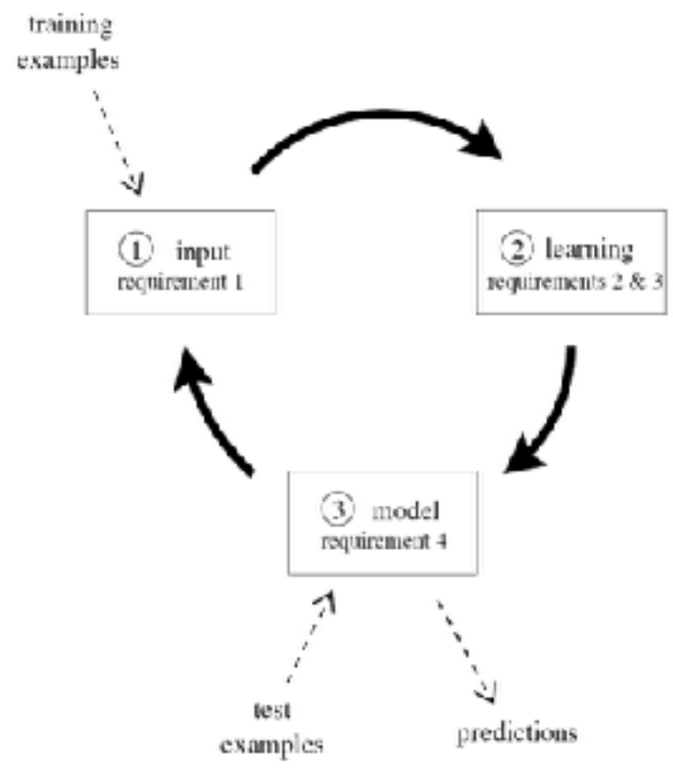


MOA: the bird

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

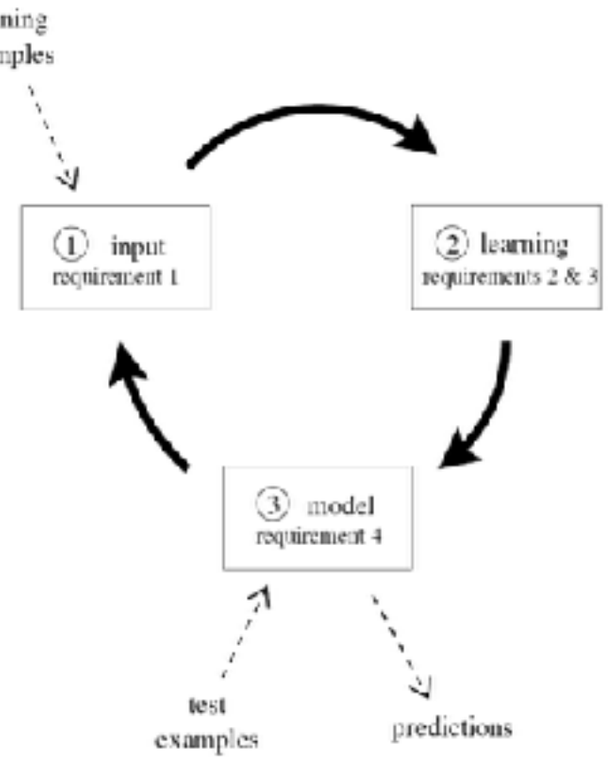
STREAM SETTING

- 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

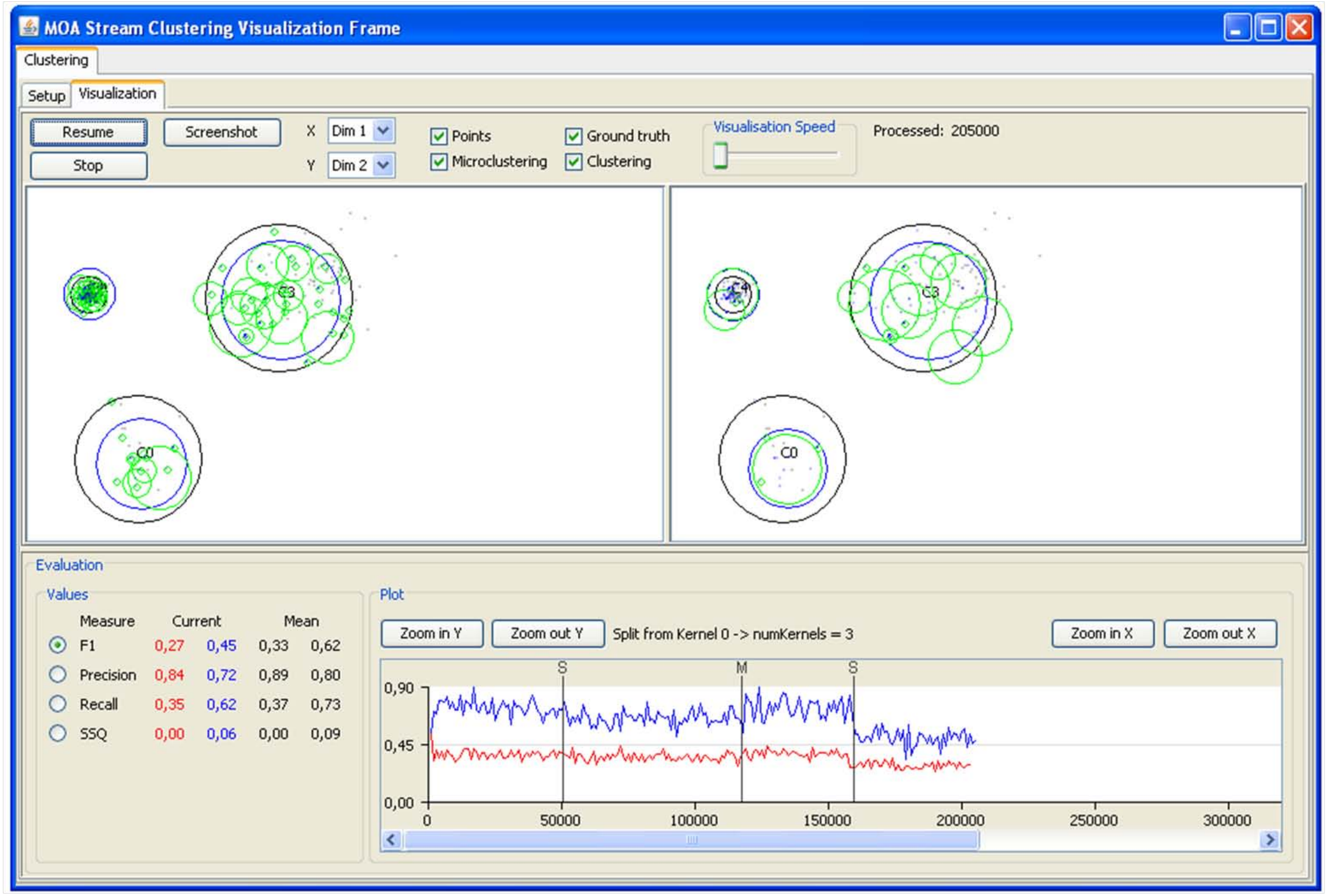


STREAM EVALUATION

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



GUI


MOA Stream Clustering Visualization Frame
Clustering
Setup
Visualization
Resume
Stop
Screenshot
X
Dim 1
Y
Dim 2
Points
Microclustering
Ground truth
Clustering
Visualisation Speed
Processed: 205000
C3
C0
C4
C3
C0
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
S
M
S
0,90
0,45
0,00
0
50000
100000
150000
200000
250000
300000

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

Data Streams

Big Data & Real Time