Real World Issues in Supervised Classification for data stream

European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (2013) […]
Workshop "**Real-World Challenges for Data Stream Mining**" […]

Vincent Lemaire
http://perso.rd.francetelecom.fr/lemaire/

Statistical learning provides numerous algorithms to build predictive models on past observations. These techniques proved their ability to deal with large scale realistic problems. However, new domains generate more and more data. This large amount of data (the buzz "big data") can be dealt with using batch algorithms (parallelized . . .) if the paradigm to store the data is realistic. But sometimes data are only visible once and need to be processed sequentially. These volatile data, known as data streams, come from telecommunication network management, social network, web mining, to name a few. The challenge is to build new algorithms able to learn under these constraints. The aim of this presentation will be to present several studies and research topics at Orange focusing on "supervised classification in data streams", with the idea to stimulate a discussion on "the real issues".

bibtex:

```
@MISC{Lemaire2013,
	author = {Lemaire, Vincent},
	title = {Real World Issues in Supervised Classification for data stream},
	year = {2013},
	note = {Slides of a talk given at ECML 2013  - workshop RealStream, September 2013},
	url = {http://perso.rd.francetelecom.fr/lemaire/publis/ECML-Pragues-vf-2013.pdf}
}
```

Outline

1. Classification using Big Data versus Classification on Stream Mining
2. Different forms of learning
3. Stream: what changes?
4. Requirements for a good algorithm
5. Taxonomy of classifier for data stream
6. Leading classifiers
7. Concept drift
8. Evaluation
9. The two streams?
10. A labeled stream?
11. Links with active learning
12. Alternative problem settings?
13. Just to do a small provocation… ☺

BIG DATA VERSUS STREAM MINING

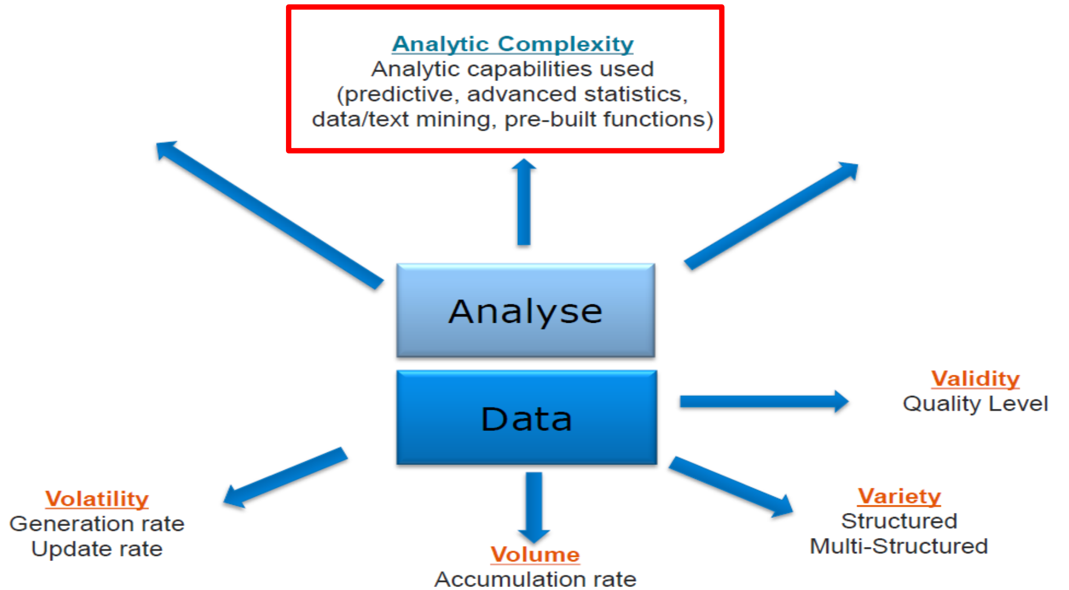
Big Data – what does that mean?

	From…	…to
Volume	Internal (terabyte)	External (petabyte)
Velocity	Batch	Real time
Variety	Structured	Unstructured
Visualization	Static	Dynamic
Value creation	Reporting and compliance	Strategic asset / Competitive advantage

McKinsey & Company | 3

Big Data ?

- Big Data = Large scale (data volume) analytics (lines & columns)
- Big Data = Emerging new data types (new multi-structured data types: web logs, sensor networks, social networks)
- Big Data = New (non-SQL) analytics (new Frameworks that provide parallel processing)



Source: BI Research 2012

Big Data Analytics ?

- Big Data Analytics : Extracting Meaningful and Actionable Data from a Massive Source
- Let’s avoid
  - Triviality, Tautology: a series of self-reinforcing statements that cannot be disproved because they depend on the assumption that they are already correct
  - Thinking that noise is an information
- Let’s try to have
  - Translation: capacity to transfer in concrete terms the discovery (actionable information)
  - TTM: Time To Market, ability to have quickly information on every customers (Who, What, Where, When)

Data Analytics - Orange Labs – team PROF

2000 2005 2010 2012

* Needs for datamining at France Telecom : a great variety of contexts

• Growing demand both in data variety and data volume

* Studies diversification, Clients outside Orange

2000 – 2005 :
Optimal Data Preparation
Automatic variable selection
Supervised Classification Models

CAP –Customer Analysis Platform
With Score Team (C. Riwan)
(CDRs from landlines)
Khiops – FT Scoring tool
...

2006 – 2010 : Scoring Factory & Khiops
Extending to various modelisation types
Extending volume capacities
Porting to Linux + 64bits

PAC PUB – moving to Hadoop
With Orange Portail France
(Usage Data from DoubleClic)

2011 – 2013 :
Multi-tables
Automatic variable construction
Co-clustering, Clustering,
Stream, …

...

Strengths

- Generic tool
- No user parameter
- Robustness
- Fineness
- Interpretability
- Efficient

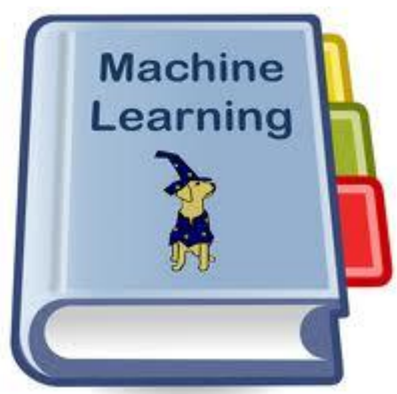
→ STREAM MINING IS REQUIRED… SOMETIMES

and…

Do not make the confusion!

Between Online Learning

and Online Deployment



A lot of advantages and drawback for both – but offline learning used 99% of the time

Machine Learning: What are the pros and cons of offline vs. online learning?

Try to find answers to:
(which is which)

- Computationally much faster and more space efficient
- Usually easier to implement
- A more general framework.
- More difficult to maintain in production.
- More difficult to evaluate online
- Usually more difficult to get "right".
- More difficult to evaluate in an offline setting, too.
- Faster and cheaper
- …

Outline

1. Classification using Big Data versus Classification on Stream Mining
2. Different forms of learning
3. Stream: what changes?
4. Requirements for a good algorithm
5. Taxonomy of classifier for data stream
6. Leading classifiers
7. Concept drift
8. Evaluation
9. The two streams?
10. A labeled stream?
11. Links with active learning
12. Alternative problem settings?
13. Just to do a small provocation… ☺

DIFFERENT FORMS OF LEARNING

Offline learning

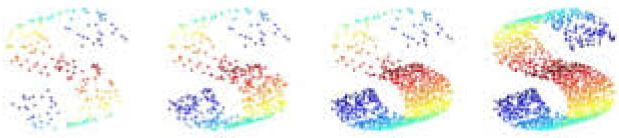
- Offline learning:
  - Examples are collected before learning starts
  - Distinct learning phase
    - It is then called “one shot learning”
  - Learning could be done:
    - Example by example
    - In batches of examples

Sequential learning

- Problems in batch learning
  - hard to control and easy to diverge
  - Can't deal with data being available over time
  - The algorithm needs to remember all training data
- Sequential learning properties
  - Examples become available over time
  - Examples are received one-by-one or by chunks
  - The model discards examples after processing them
  - Doesn't require retraining each time an example comes
  - The algorithm stops and returns a unique hypothesis when examples end, or when a certain condition is verified.
  - The algorithm stores only model's parameters (e.g. weights)
  - Two distinct phases: learning and operation

Incremental learning

- properties
  - Inherits all properties of online learning
    - i.e. online learning is always incremental
  - But it is more general
    - It can be also offline
    - It can handle batches of data
  - Shares with sequential and online learning the obligation:
    - To Process new data and discard them later
    - Learn new information and retain old one's
      - →stability-plasticity dilemma



Online learning

- Properties
  - Receives examples one-by-one
  - discards the example after processing it.
  - Produce a hypothesis after each example is processed
    - i.e. produces a series of hypotheses
  - No distinct phases for learning and operation
    - i.e. produced hypotheses can be used in classification
  - Allowed to store other parameters besides model parameters (e.g. learning rate)
  - Is a real time system
    - Constraints: time, memory, …
    - What is affected: hypotheses prediction accuracy
  - Can never stop

Anytime learning

Focus today - Supervised classifier

- Try to find answers to:
  - Can the examples be stored in memory?
  - Which is the availability of the examples: any presents? In stream ? Visible only once?
  - Is the concept stationary?
  - Does the algorithm have to be anytime?
  - What is the available time to update the model?
  - …

- The answers to these questions will give indications to select the algorithms adapted to the situation and to know if one need an incremental algorithm, even a specific algorithm for data stream.

Outline

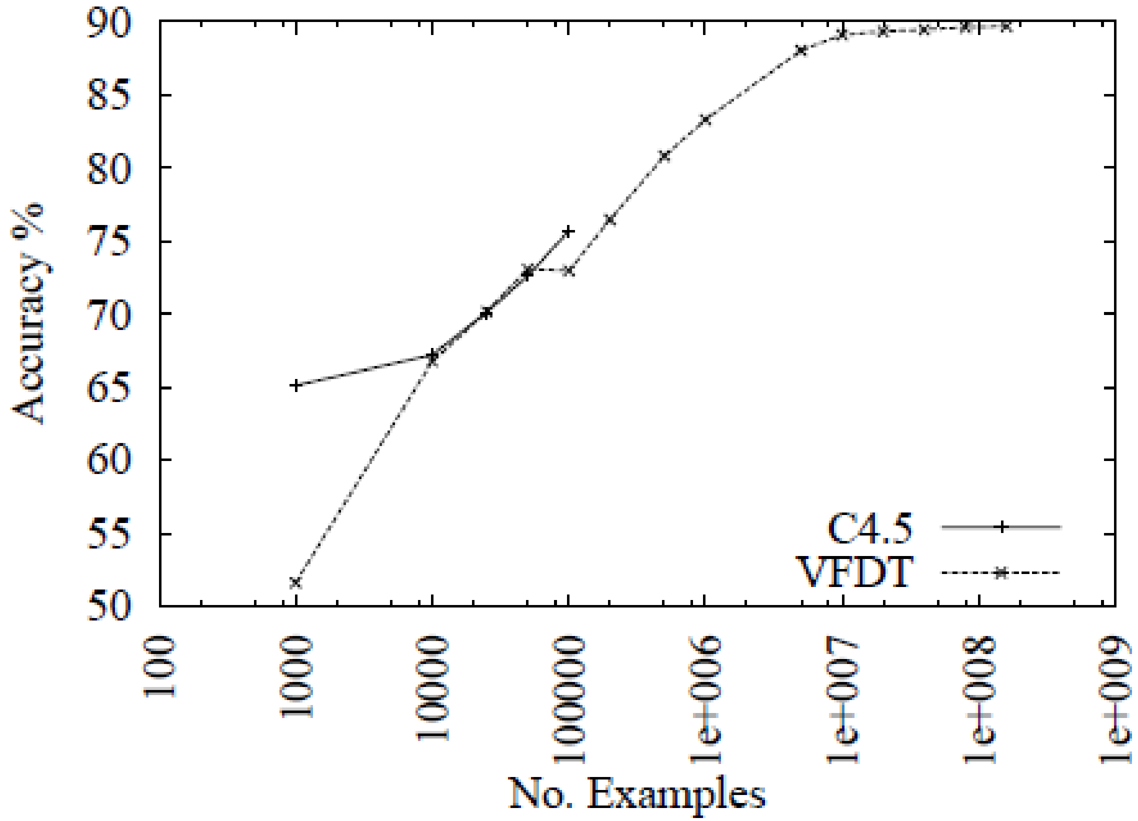
1. Classification using Big Data versus Classification on Stream Mining
2. Different forms of learning
3. Stream: what changes?
4. Requirements for a good algorithm
5. Taxonomy of classifier for data stream
6. Leading classifiers
7. Concept drift
8. Evaluation
9. The two streams?
10. A labeled stream?
11. Links with active learning
12. Alternative problem settings?
13. Just to do a small provocation… ☺

STREAM: WHAT CHANGES?

Stream: what changes?

- Properties
  - Receives examples one-by-one
  - discards the example after processing it.
  - Produce a hypothesis after each example is processed
    - i.e. produces a series of hypotheses
  - No distinct phases for learning and operation
    - i.e. produced hypotheses can be used in classification
  - Allowed to store other parameters than model parameters (e.g. learning rate)
  - Is a real time system
    - Constraints: time, memory, …
    - What is affected: hypotheses prediction accuracy
  - Can never stop
  - No i. i. d

Why not use the classic algorithms?



Domingos, P., & Hulten, G. (2000). Mining high-speed data streams. *SIGKDD*

Outline

REQUIREMENTS

Properties of a efficient algorithm

- low and constant duration to learn from the examples;
- read only once the examples in their order of arrival;
- use of a quantity of memory fixed “a priori;”
- production of a model close to the “offline model”
- anytime
- concept drift management

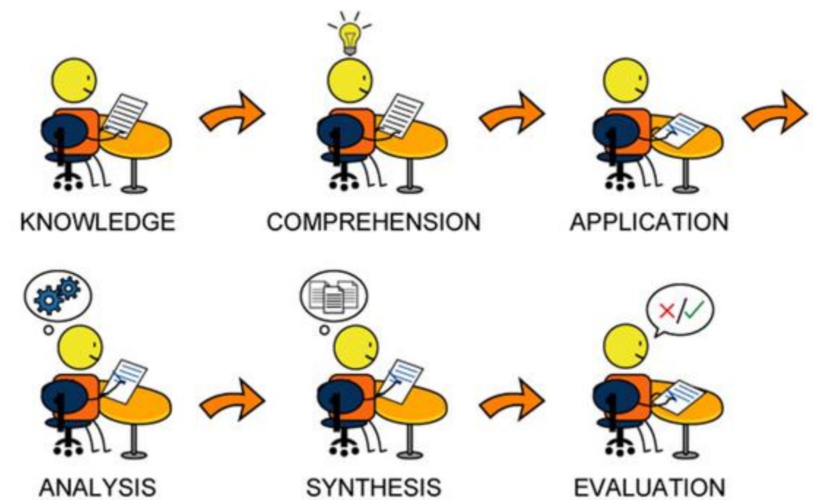
(0) Domingos, P. et G. Hulten (2001). Catching up with the data : Research issues in mining data streams. In Workshop on Research Issues in Data Mining and Knowledge Discovery.
(1) Fayyad, U. M., G. Piatetsky-Shapiro, P. Smyth, et R. Uthurusamy (1996). Advances in Knowledge Discovery and Data Mining. Menlo Park, CA, USA : American Association for Artificial Intelligence
(2) Hulten, G., L. Spencer, et P. Domingos (2001). Mining time-changing data streams. In Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 97–106. ACM New York, NY, USA.
(3) Stonebraker, M., U. Çetintemel, et S. Zdonik (2005). The 8 requirements of real-time stream processing. ACM SIGMOD Record 34(4), 42–47.

Outline

1. Classification using Big Data versus Classification on Stream Mining
2. Different forms of learning
3. Stream: what changes?
4. Requirements for a good algorithm
5. Taxonomy of classifier for data stream
6. Leading classifiers
7. Concept drift
8. Evaluation
9. The two streams?
10. A labeled stream?
11. Links with active learning
12. Alternative problem settings?
13. Just to do a small provocation… ☺

Vincent Lemaire - (c) 2013 - Orange Labs

TAXONOMY


KNOWLEDGE
COMPREHENSION
APPLICATION
ANALYSIS
SYNTHESIS
EVALUATION

Taxonomy (tentative)

full example memory Store *all* examples
- allows for efficient restructuring
- good accuracy
- huge storage needed

Examples: ID5, ID5R, ITI

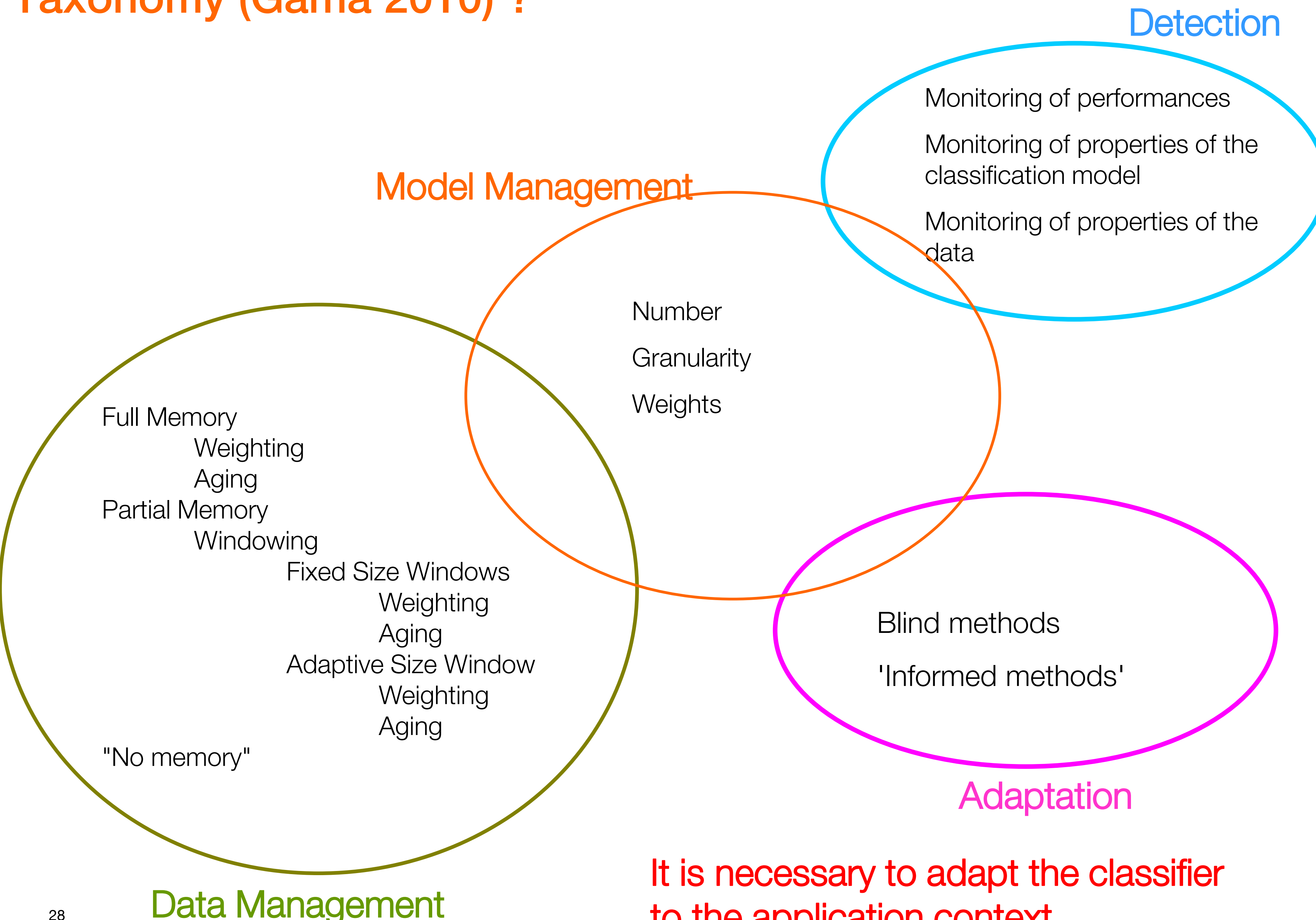
no example memory Only store statistical information in the nodes
- loss of accuracy (depending on the information stored or again huge storage needed)
- relatively low storage space

Examples: ID4

partial example memory Only store *selected* examples
- trade of between storage space and accuracy

Examples: FLORA, AQ-PM

Taxonomy (Gama 2010) ?



It is necessary to adapt the classifier to the application context

Outline

1. Classification using Big Data versus Classification on Stream Mining
2. Different forms of learning
3. Stream: what changes?
4. Requirements for a good algorithm
5. Taxonomy of classifier for data stream
6. Leading classifiers
7. Concept drift
8. Evaluation
9. The two streams?
10. A labeled stream?
11. Links with active learning
12. Alternative problem settings?
13. Just to do a small provocation… ☺

LEADING CLASSIFIERS

Incremental Algorithm (no stream)

- Decision Tree
  - ID4 (Schlimmer - ML'86)
  - ID5/ITI (Utgoff – ML'97)
  - SPRINT (Shaffer - VLDB'96)
  - …
- Naive Bayes
  - Incremental (for the standard NB)
  - Learn fastly with a low variance (Domingos – ML'97)
  - Can be combined with decision tree: NBTree (Kohavi – KDD'96)

Incremental Algorithm (no stream)

- Neural Networks
  - IOLIN (Cohen - TDM’04)
  - learn++ (Polikar - IJCNN’02),…
- Support Vector Machine
  - TSVM (Transductive SVM – Klinkenberg IJCAI’01),
  - PSVM (Proximal SVM – Mangasarian KDD’01),…
  - LASVM (Bordes 2005)
- Rules based systems
  - AQ15 (Michalski - AAAI’86), AQ-PM (Maloof/Michalski - ML’00)
  - STAGGER (Schlimmer - ML’86)
  - FLORA (Widmer - ML’96)

Incremental Algorithm (for stream)

- Rules
  - FACIL (Ferrer-Troyano – SAC'04,05,06)
- Ensemble
  - SEA (Street - KDD'01) based on C4.5
- K-nn
  - ANNCAD (Law – LNCS'05).
  - IBLS-Stream (Shaker et al – Evolving Systems" journal 2012)
- SVM
  - CVM (Tsang – JMLR'06)

Incremental Algorithm (for stream)

- Decision Tree – the only ones used ?
  - Domingos : VFDT (KDD’00), CVFDT (KDD’01)
  - Gama : VFDTc (KDD’03), UFFT (SAC’04)
  - Kirkby : Ensemble d’Hoeffding Trees (KDD’09)
  - del Campo-Avila : IADEM (LNCS’06)

Outline

1. Classification using Big Data versus Classification on Stream Mining
2. Different forms of learning
3. Stream: what changes?
4. Requirements for a good algorithm
5. Taxonomy of classifier for data stream
6. Leading classifiers
7. Concept drift
8. Evaluation
9. The two streams?
10. A labeled stream?
11. Links with active learning
12. Alternative problem settings?
13. Just to do a small provocation… ☺

CONCEPT DRIFT

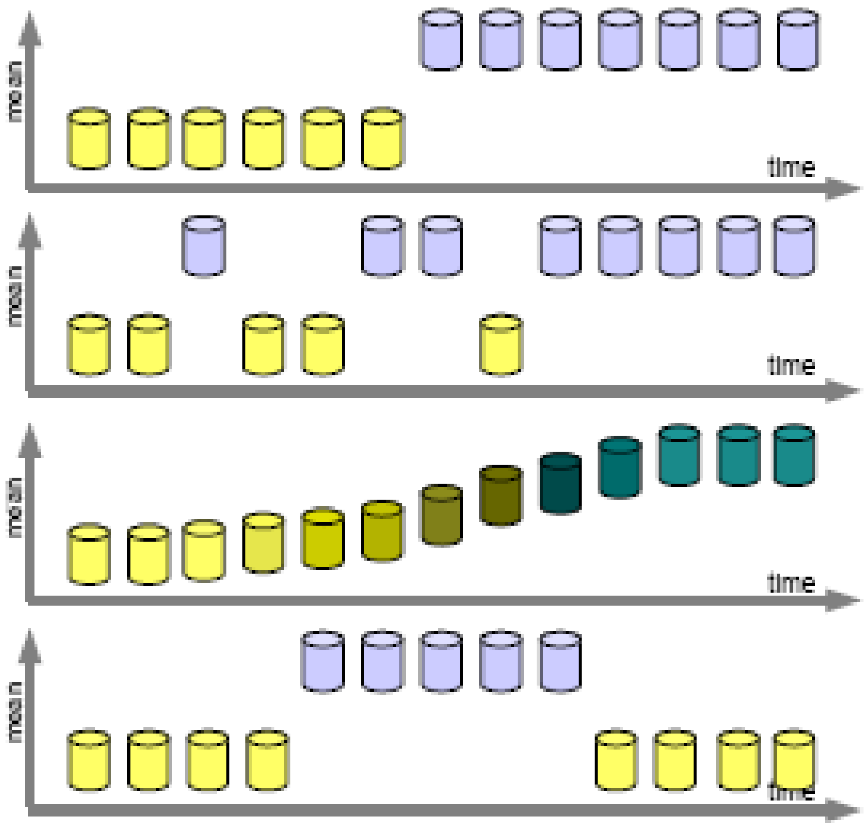
Concept drift

Concept shift

Gradual drift

Incremental shift

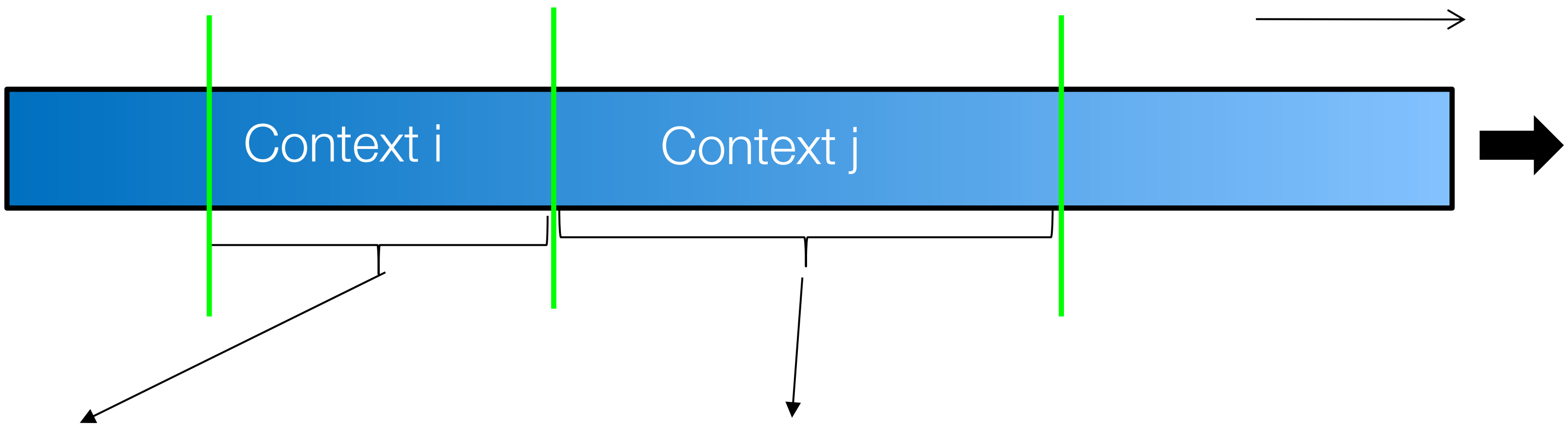
Reoccurring contexts



Learning under Concept Drift: an Overview

Indrė Žliobaitė
Faculty of Mathematics and Informatics
Vilnius University, Lithuania zliobaite@gmail.com

Context...



$$P_1(x,y) = P_1(x)P_1(y|x)$$

$$P_2(x,y) = P_2(x)P_2(y|x)$$

$$P_2(x,y) = P_1(x)P_2(y|x)$$

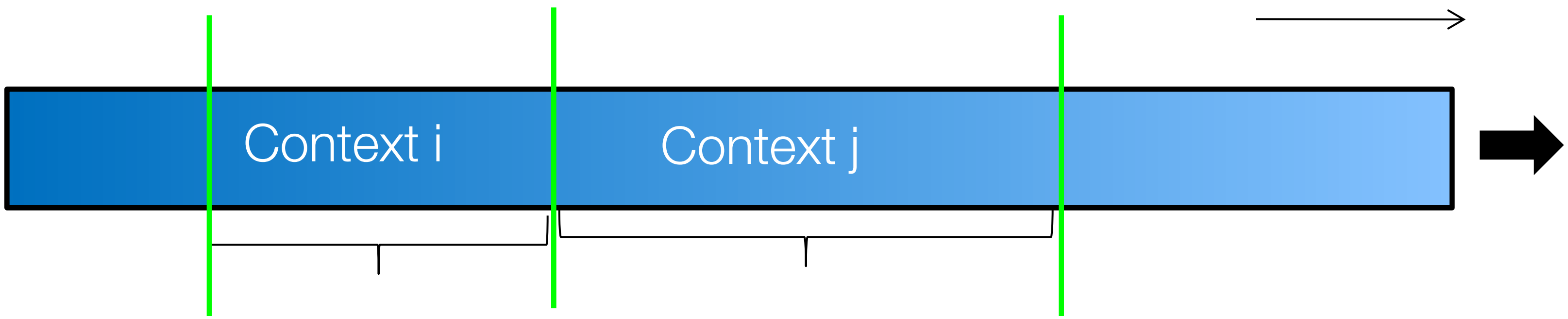
$$P_2(x,y) = P_2(x)P_1(y|x)$$

Context = Period of time without drift

Stream: sequence of context

Drift detection ⇔ ? Manage drift

Manage Drift?



- Either detect and :
  1) Retrain the model
  2) Adapt the current model
  3) Adapt statistics (summaries) on which the model is based
  4) Work with a sequence of
     - models
     - summaries
- or detect anything but train (learn) fastly
  - a single models
  - an ensemble of models)

Parameters – The devil inside

No drift assumption?

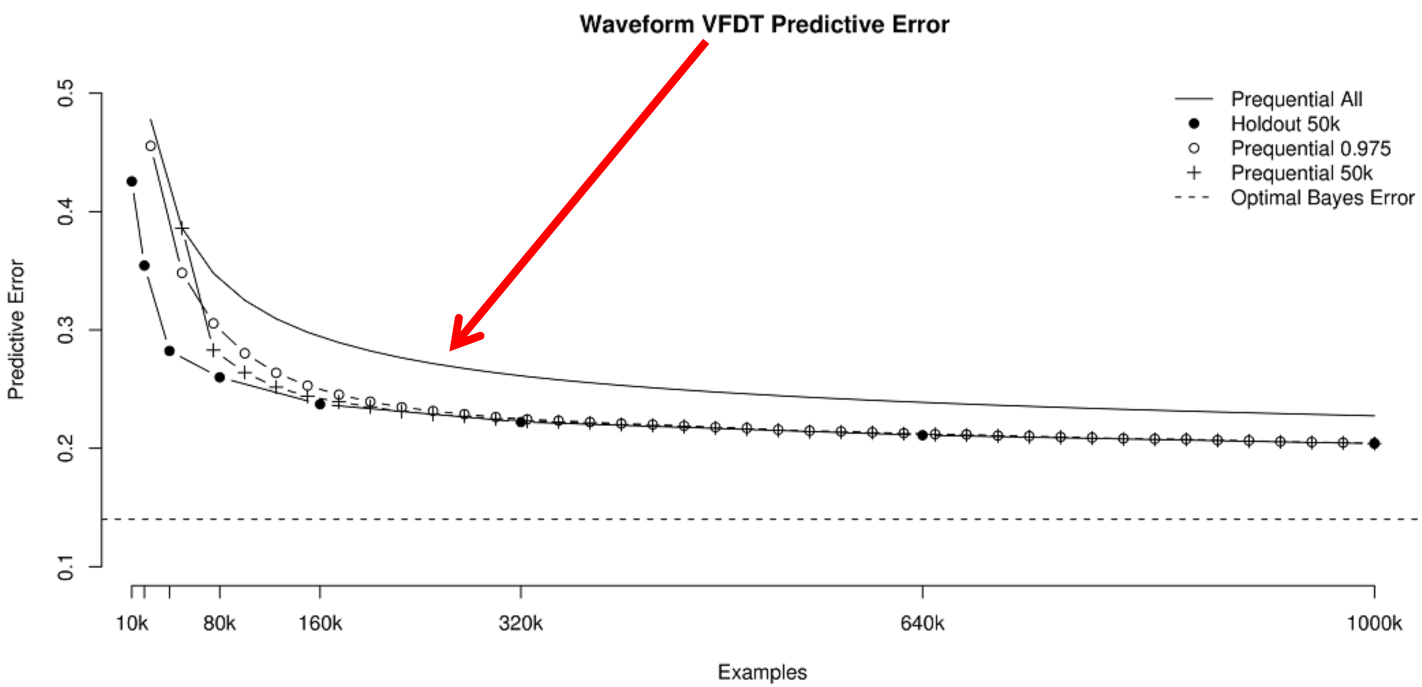
Do not use online learning !

Outline

1. Classification using Big Data versus Classification on Stream Mining
2. Different forms of learning
3. Stream: what changes?
4. Requirements for a good algorithm
5. Taxonomy of classifier for data stream
6. Leading classifiers
7. Concept drift
8. Evaluation
9. The two streams?
10. A labeled stream?
11. Links with active learning
12. Alternative problem settings?
13. Just to do a small provocation… ☺

EVALUATION

Prequential error – the only way ?



Pessimistic

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

But other ideas

- Littlestone, N. et M. Warmuth (1989). The weighted majority algorithm. 30th Annual Symposium on Foundations of Computer Science, 256–261.
  - Mystake-bound

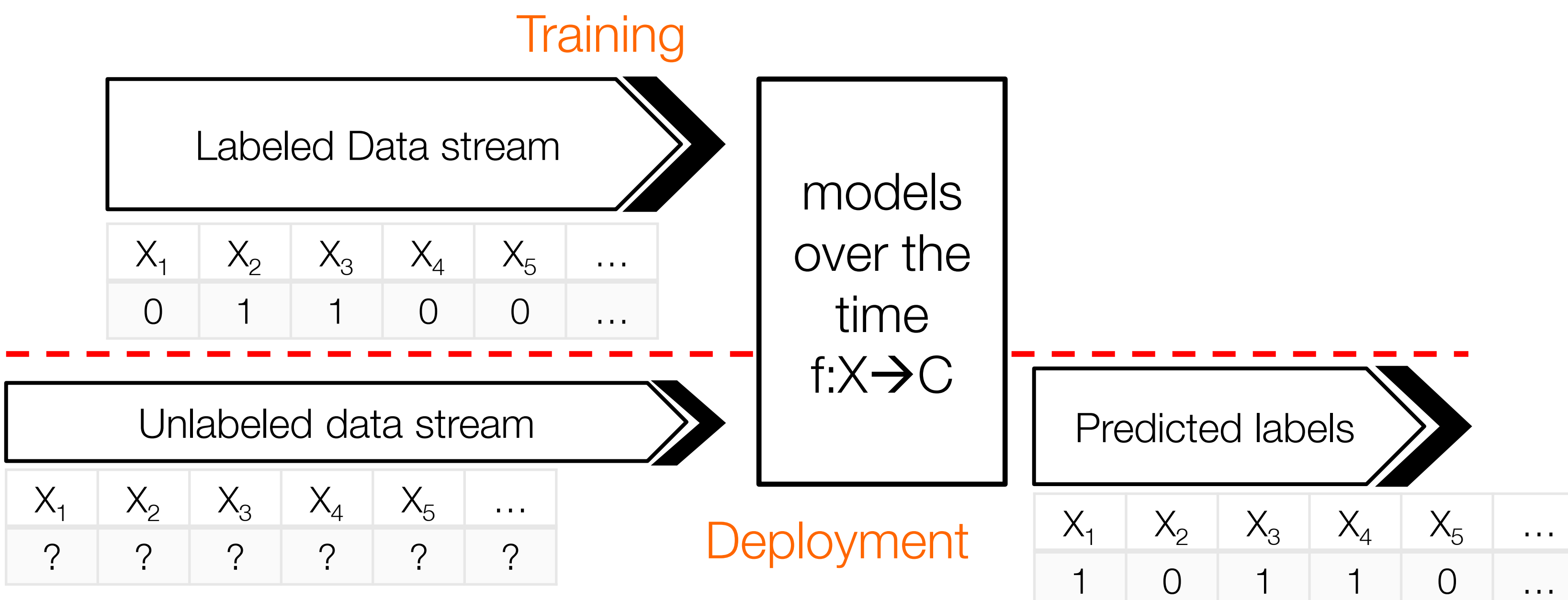
Outline

1. Classification using Big Data versus Classification on Stream Mining
2. Different forms of learning
3. Stream: what changes?
4. Requirements for a good algorithm
5. Taxonomy of classifier for data stream
6. Leading classifiers
7. Concept drift
8. Evaluation
9. The two streams?
10. A labeled stream?
11. Links with active learning
12. Alternative problem settings?
13. Just to do a small provocation… ☺

THE TWO STREAMS?

Supervised Classification

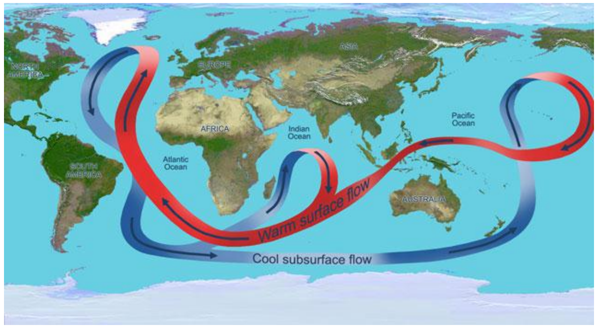
- Two streams exist
- Two drift detection have to be managed



Outline

1. Classification using Big Data versus Classification on Stream Mining
2. Different forms of learning
3. Stream: what changes?
4. Requirements for a good algorithm
5. Taxonomy of classifier for data stream
6. Leading classifiers
7. Concept drift
8. Evaluation
9. The two streams?
10. A labeled stream?
11. Links with active learning
12. Alternative problem settings?
13. Just to do a small provocation… ☺

THE LABELED STREAM?

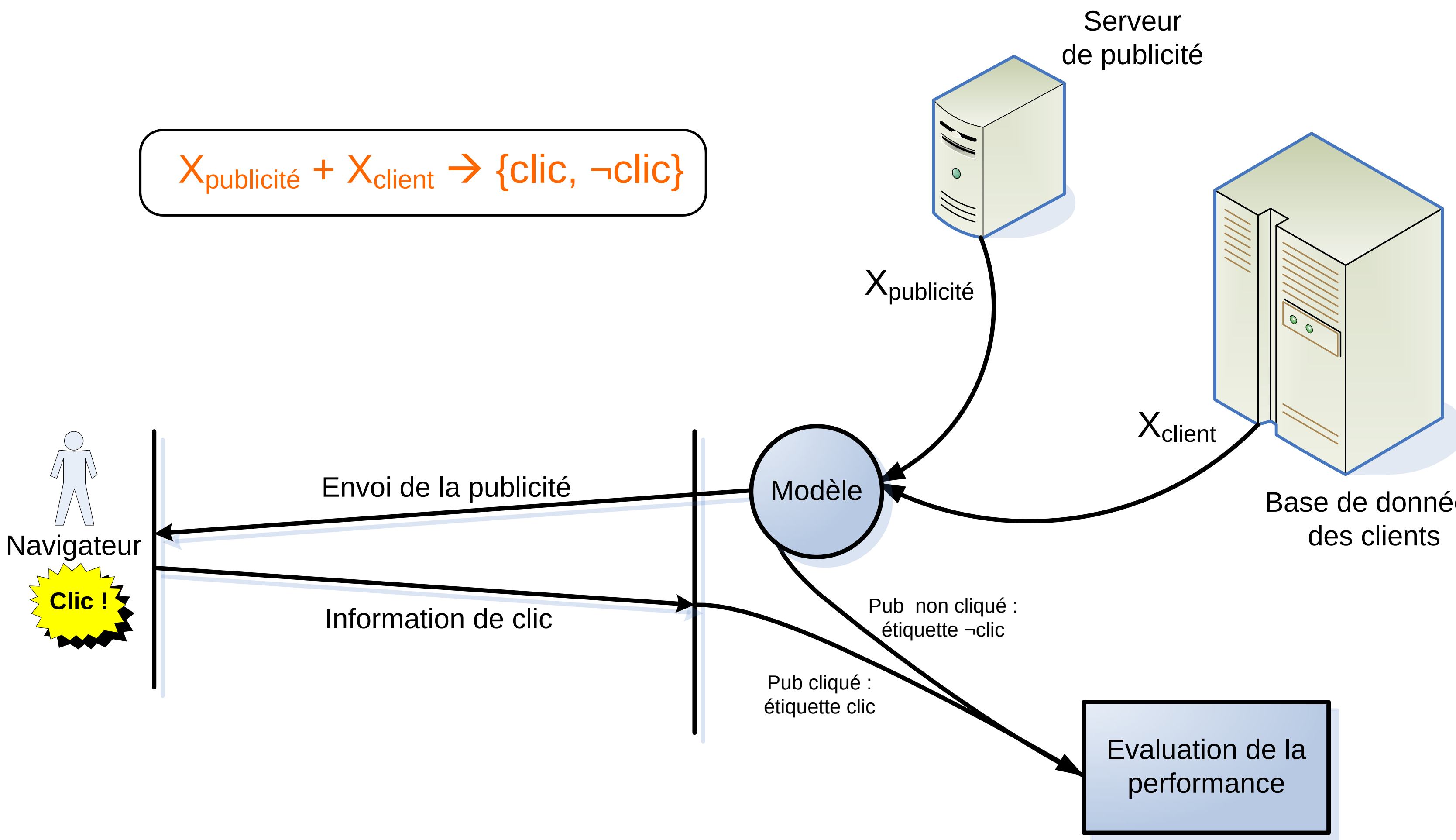


even for classic problem not easy to obtain…

Online Advertising

the problem: having the ‘label’ in a « correct » timing (and for the complete distribution)

$X_{publicité} + X_{client} \rightarrow \{clic, \neg clic\}$



Outline

LINK WITH

ACTIVE LEARNING

of course but in another talk?

Outline

1. Classification using Big Data versus Classification on Stream Mining
2. Different forms of learning
3. Stream: what changes?
4. Requirements for a good algorithm
5. Taxonomy of classifier for data stream
6. Leading classifiers
7. Concept drift
8. Evaluation
9. The two streams?
10. A labeled stream?
11. Links with active learning
12. Alternative problem settings?
13. Just to do a small provocation… ☺

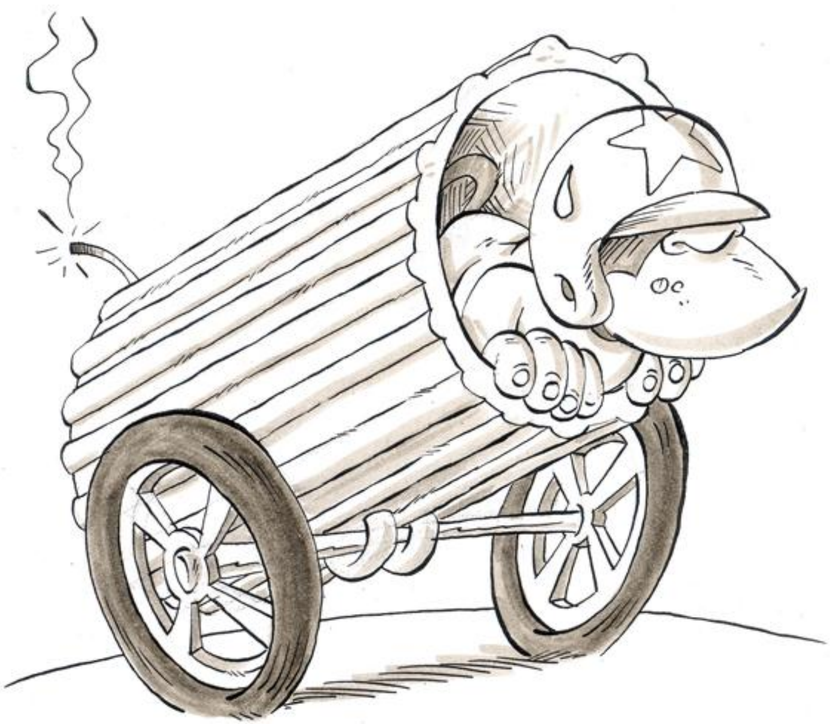
OTHER WAYS?

Alternative problem settings

Outline

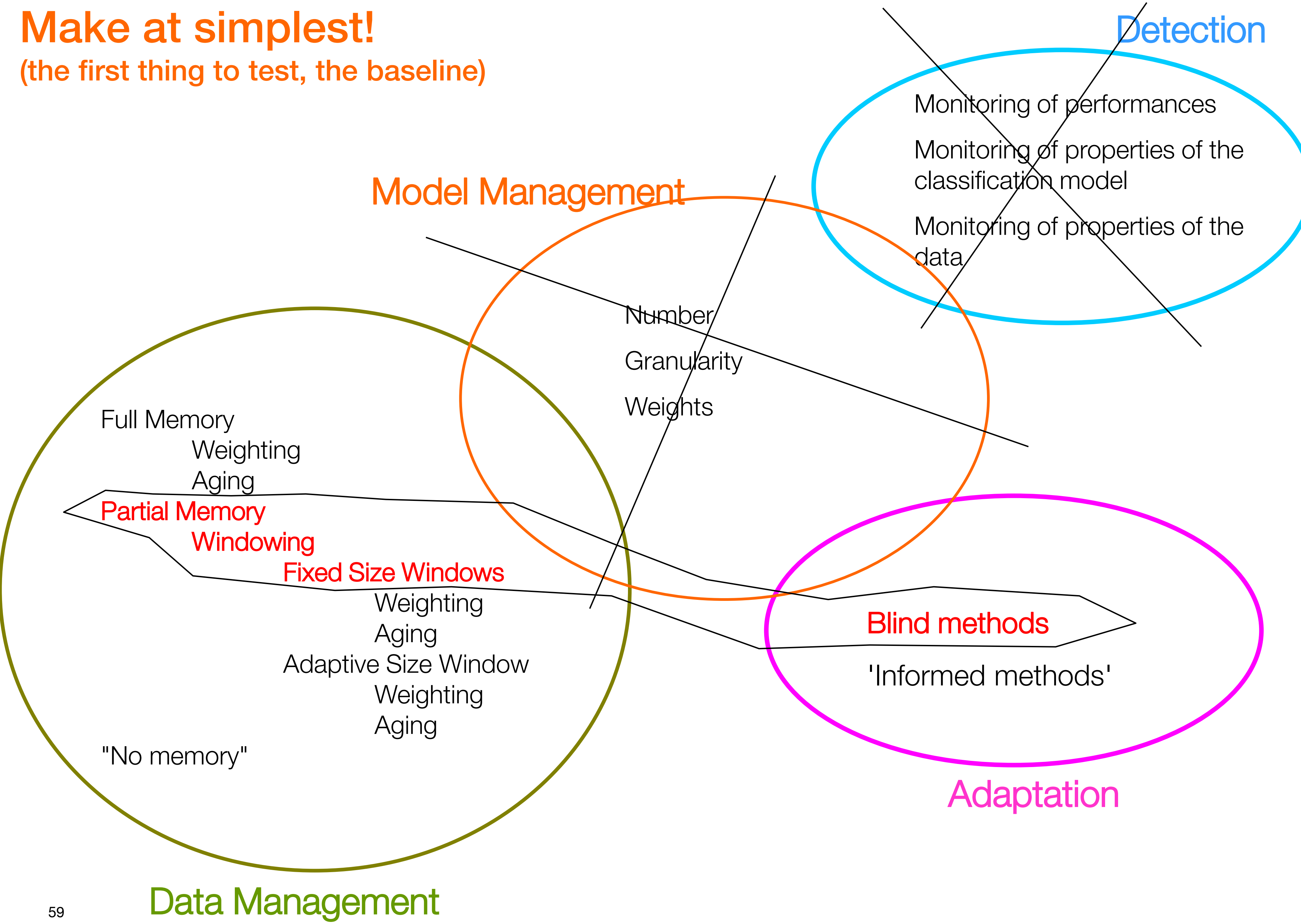
1. Classification using Big Data versus Classification on Stream Mining
2. Different forms of learning
3. Stream: what changes?
4. Requirements for a good algorithm
5. Taxonomy of classifier for data stream
6. Leading classifiers
7. Concept drift
8. Evaluation
9. The two streams?
10. A labeled stream?
11. Links with active learning
12. Alternative problem settings?
13. Just to do a small provocation… ☺

PROVOCATION



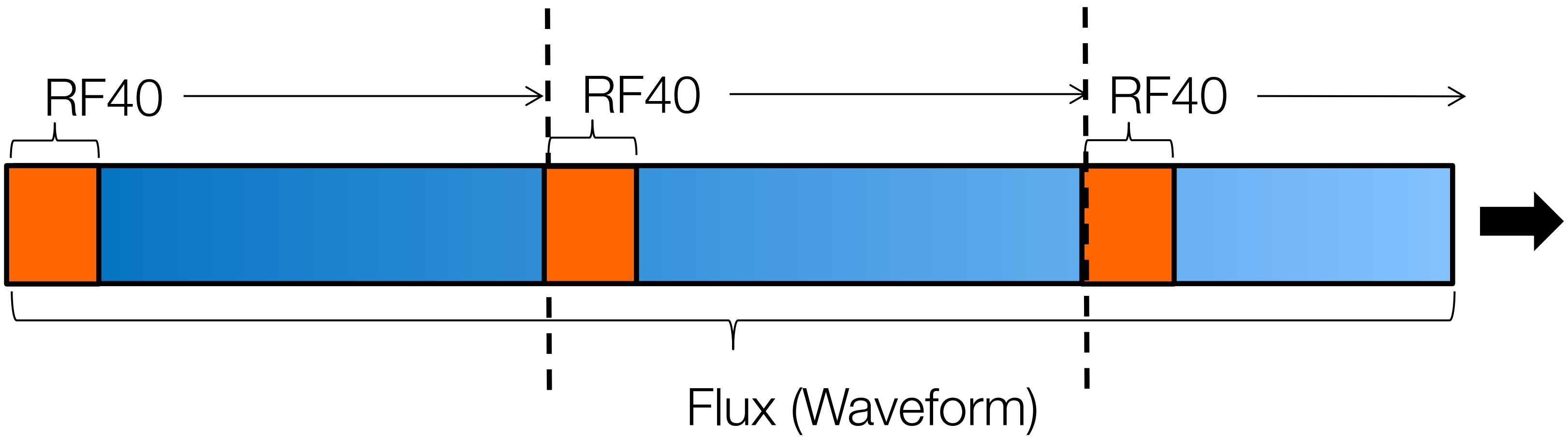
Make at simplest!

(the first thing to test, the baseline)

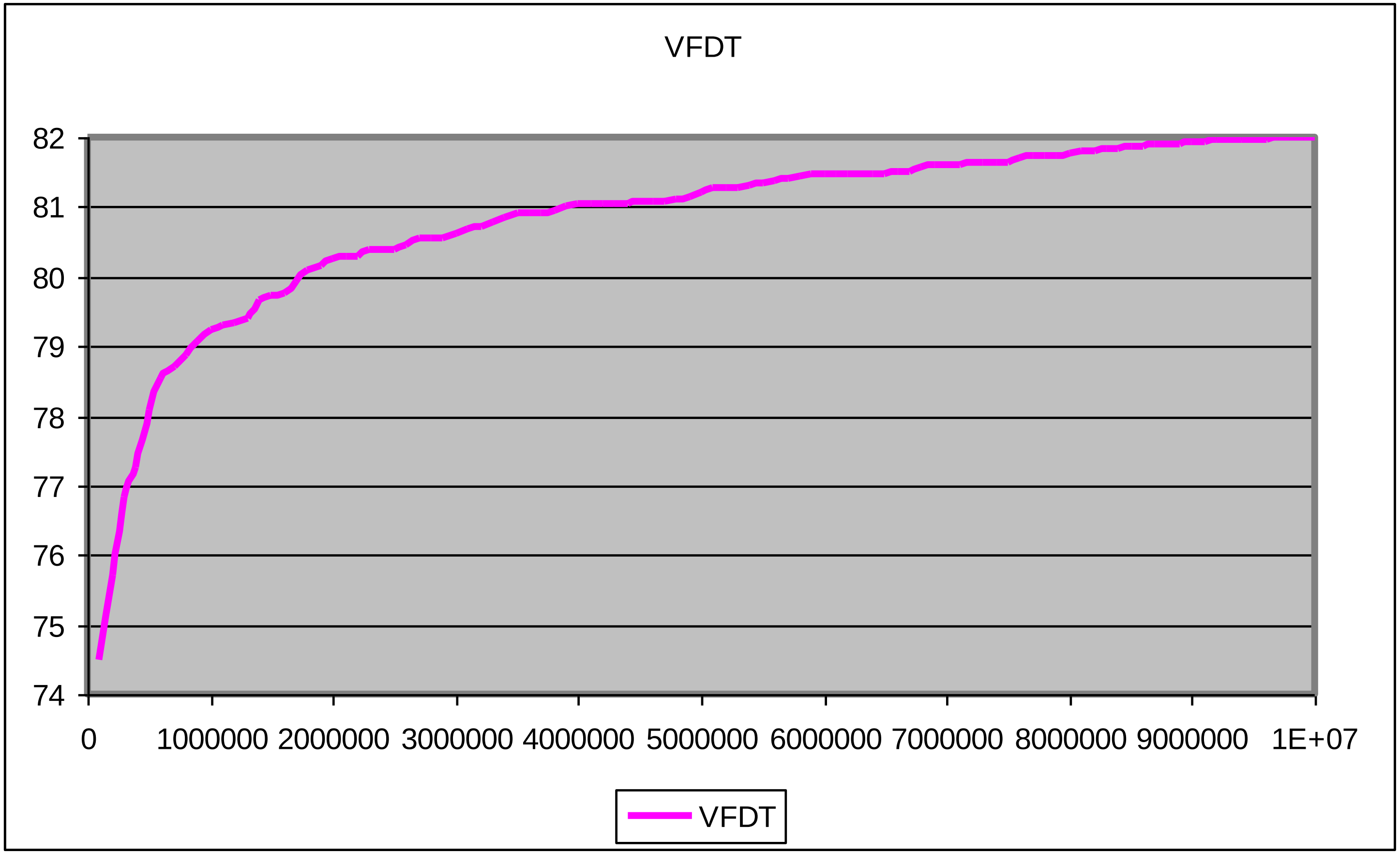


A classifier trained with few examples but often!

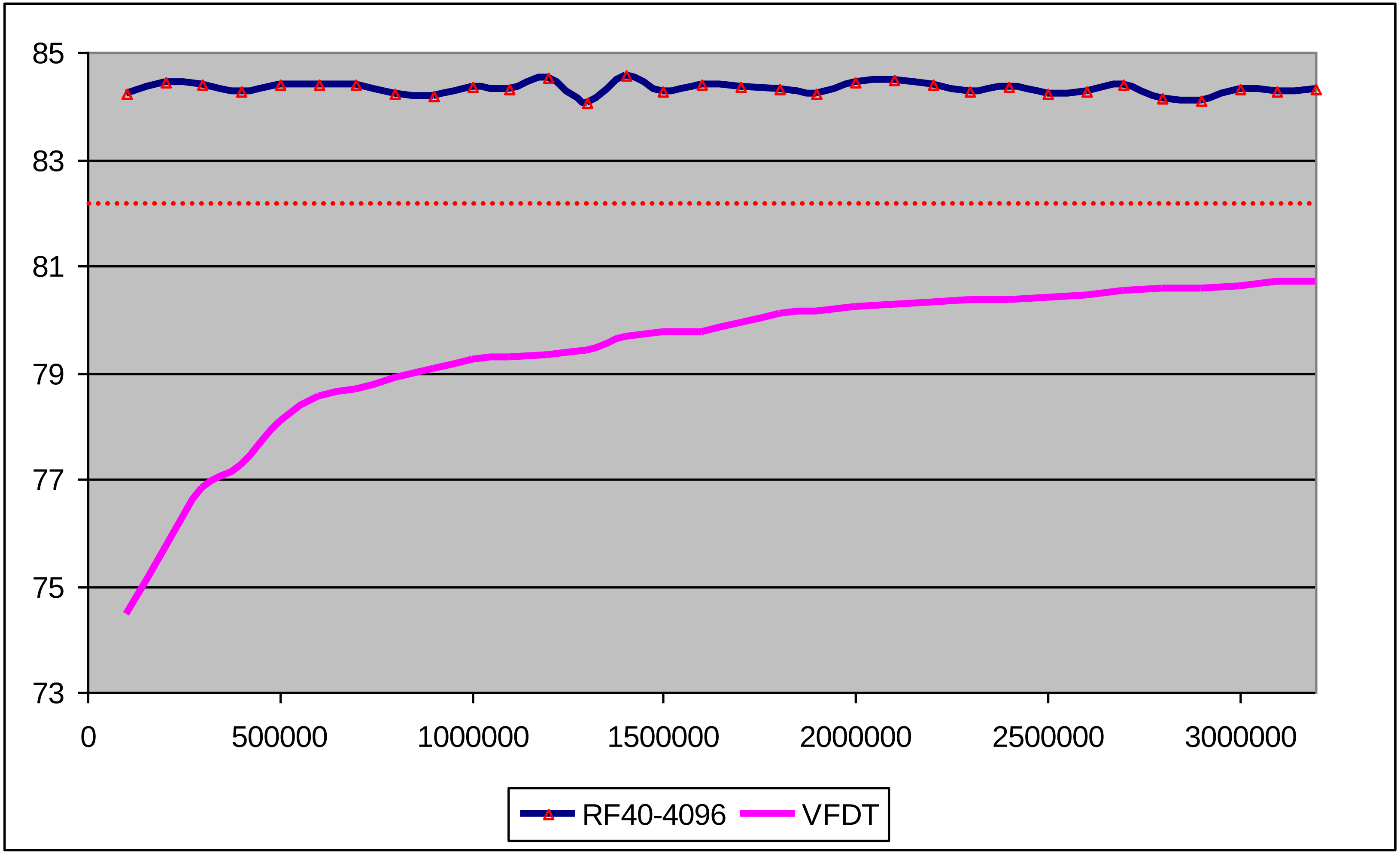
- Which classifier ?
  - a random forest (based on « "Learning with few examples: an empirical study on leading classifiers ", Christophe Salperwyck and Vincent Lemaire, in International Joint Conference on Neural Networks (IJCNN July 2011)»)
  - using 4096 examples



Waveform

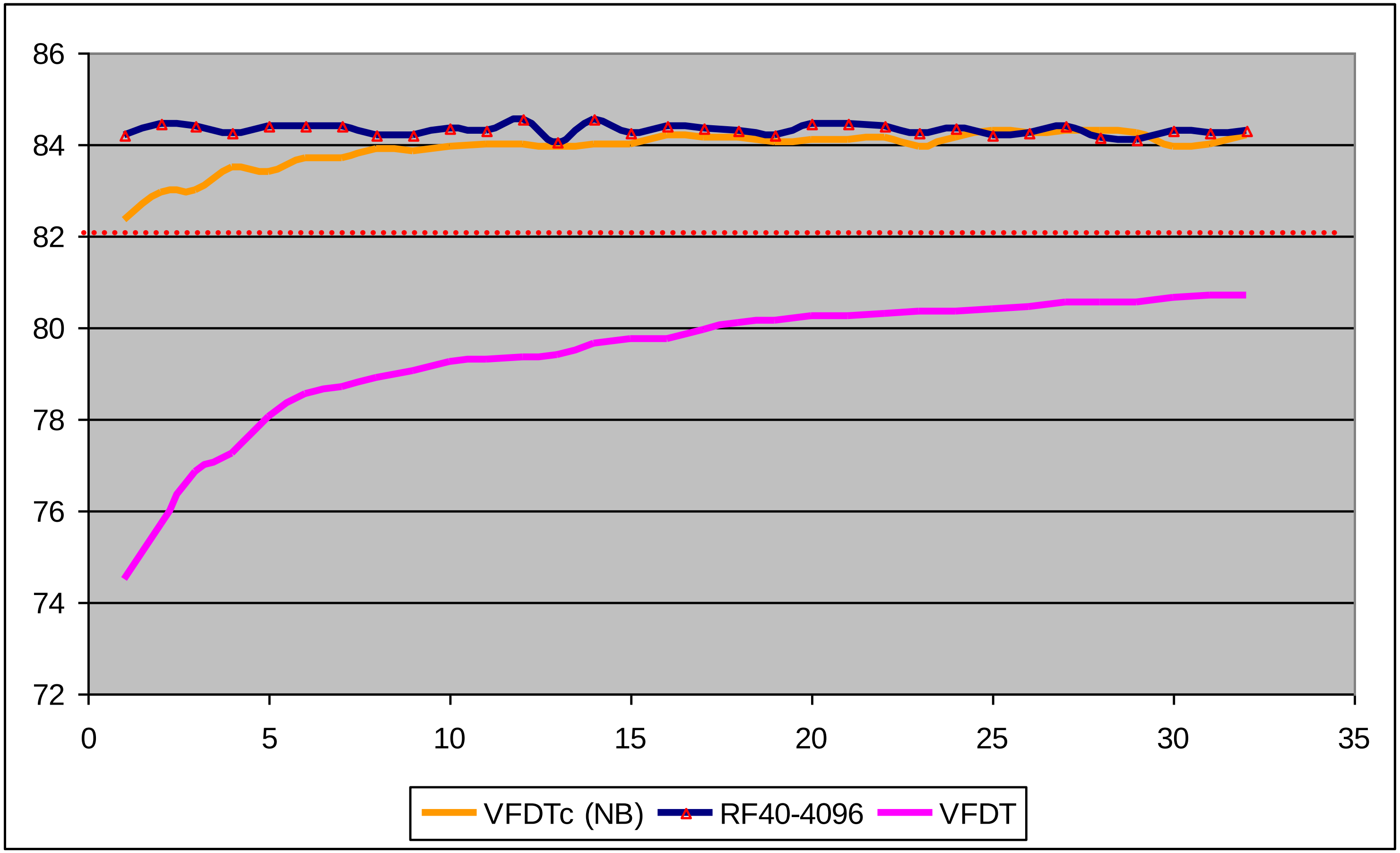

VFDT
82
81
80
79
78
77
76
75
74
0
1000000
2000000
3000000
4000000
5000000
6000000
7000000
8000000
9000000
1E+07
VFDT

Waveform



85
83
81
79
77
75
73
0
500000
1000000
1500000
2000000
2500000
3000000
RF40-4096
VFDT

Waveform



Outline

the end

Real issues - Discussion

- Different form of learning:
  - use online or very incremental algorithm only if needed
- Stream: the changes for a good algorithm
  - deal directly the tradeoff between memory used / precision / complexity
- Leading classifiers
  - tries to used fast learner
- Concept drift
  - no stream mining without the concept drift ‘management’
- Evaluation
  - use robust model and model selection (no validation) and prequential error
- The two streams?
  - not so easy to have a correct “labeled” stream, two different drifts have to be managed
- Others ways to pose the problem
  - do not forget
- Just to do a small provocation…
  - a lot of papers but not so many real applications – the parameters are the devil (and the life expectancy of the “system”)