

PAKDD 2011

24 - 27 May 2011, Shenzhen, China



Handling Concept Drift: Importance, Challenges & Solutions

A. Bifet, J. Gama, M. Pechenizkiy, I. Žliobaitė



THE UNIVERSITY OF
WAIKATO
Te Whare Wānanga o Waikato



TU/e

Technische Universiteit
Eindhoven
University of Technology



**Bournemouth
University**

PAKDD-2011 Tutorial
May 27, Shenzhen, China

<http://www.cs.waikato.ac.nz/~abifet/PAKDD2011/>

Motivation for the Tutorial

- in the real world data
 - more and more data is organized as **data streams**
 - data comes from complex environment, and it is **evolving** over time
 - **concept drift** = underlying distribution of data is changing
- attention to handling concept drift is increasing, since predictions become less accurate over time
- to prevent that, **learning models need to adapt** to changes quickly and accurately

Objectives of the Tutorial

- many approaches to handle drifts in a wide range of applications have been proposed
- the tutorial aims to provide an integrated view **of the adaptive learning methods and application needs**

Outline

Block 1: PROBLEM

Block 2: TECHNIQUES

Block 3: EVALUATION

Block 4: APPLICATIONS

OUTLOOK

Presenters & Schedule

Indrė Žliobaitė



14:05 – 14:50



Block 1: what concept drift is, why it needs to be handled and how

João Gama



14:50 – 15:40



Block 2: approaches for adaptive machine learning to handle concept drift

Presenters & Schedule

Albert Bifet



16:00 – 16:45



Block 3: training/evaluation of adaptive learning approaches and a software demo

Mykola Pechenizkiy



16:45 – 17:30



Block 4: challenges for handling concept drift in different applications



Block 1: introduction

The goals:

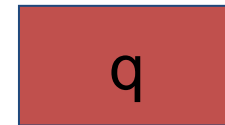
- to introduce the problem of concept drift in supervised learning
- to overview and categorize the main principles how concept drift can be (and is) handled

Example: production

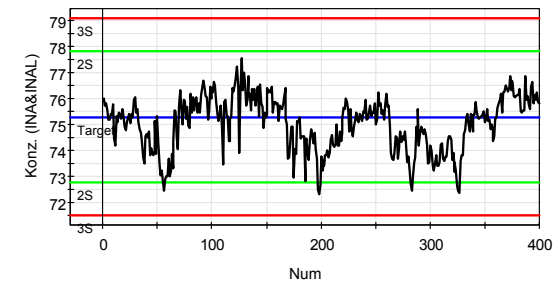
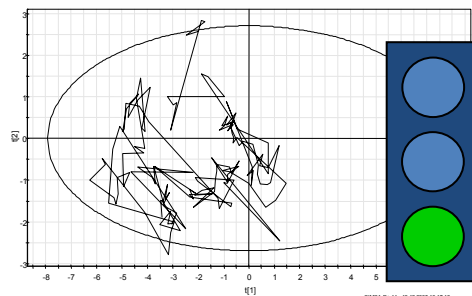


Example: production

predict quality

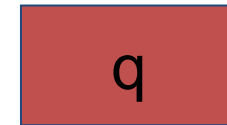


monitor

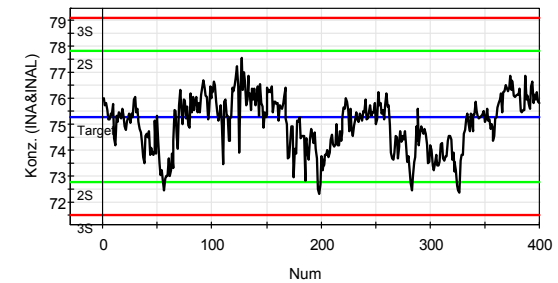
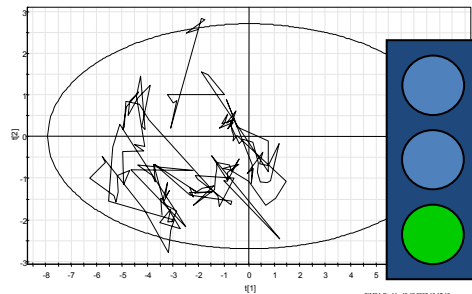


Example: production

predict quality



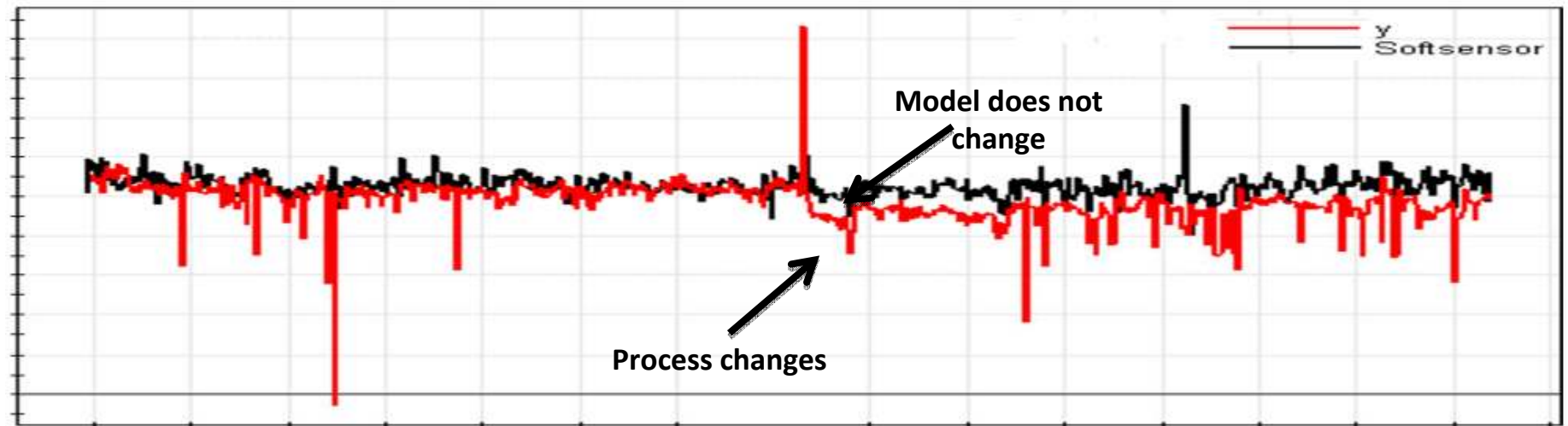
monitor



after 2-3 months
predictions get
“out of tune”



Static model



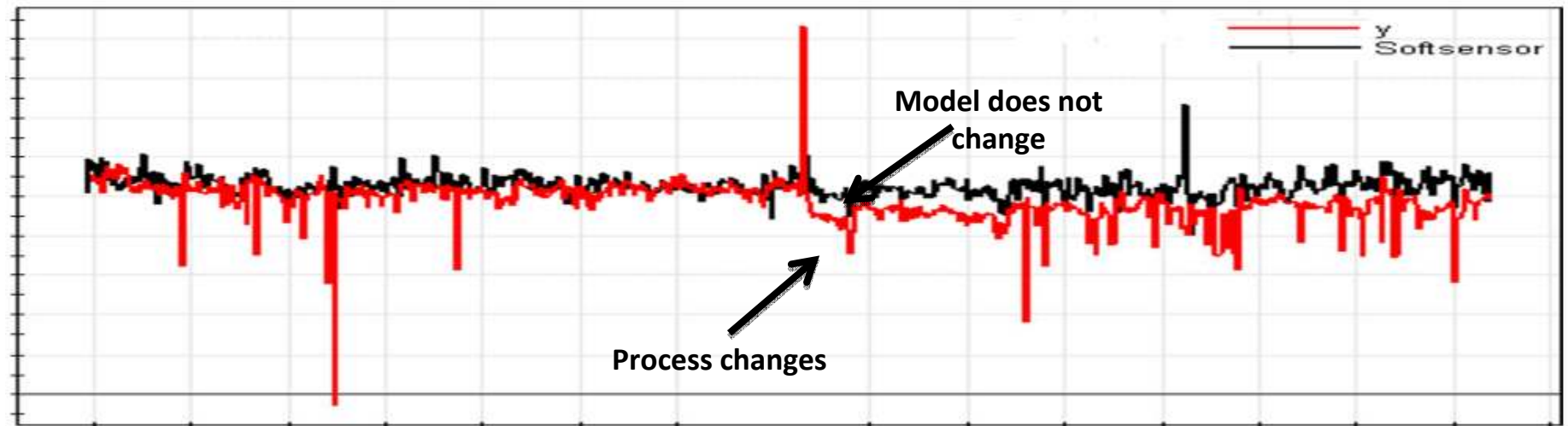
source: *Evonik Industries*



the supplier of raw material changes

a sensor breaks/ "wears off"/ is replaced

new operating crew



source: Evonik Industries

new regulations to save electricity

new production procedures



Desired Properties of a System To Handle Concept Drift

- Adapt to concept drift asap



- Distinguish noise from changes
 - robust to noise, but adaptive to changes



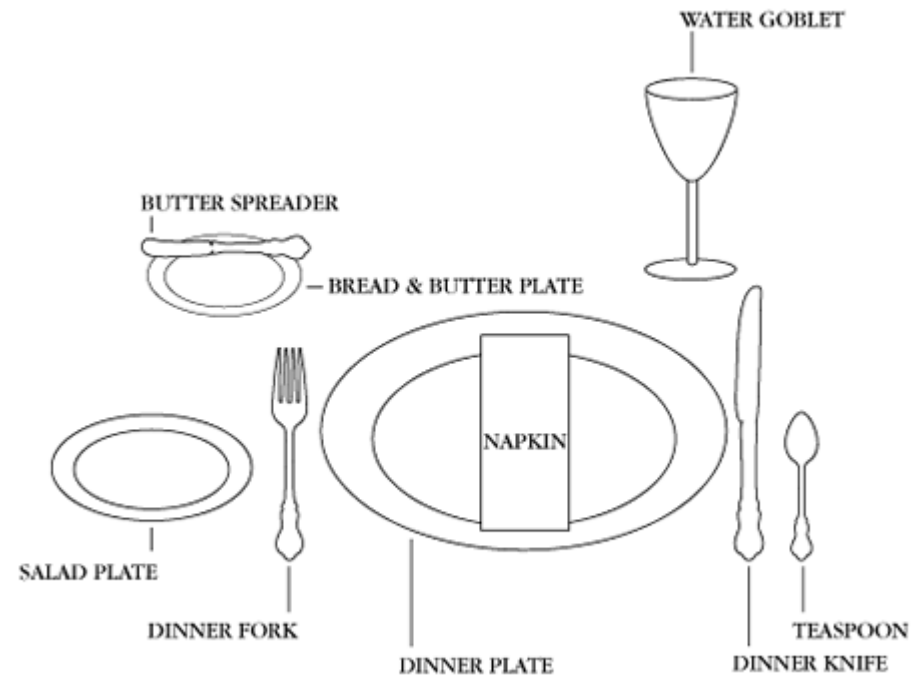
- Recognizing and reacting to reoccurring contexts



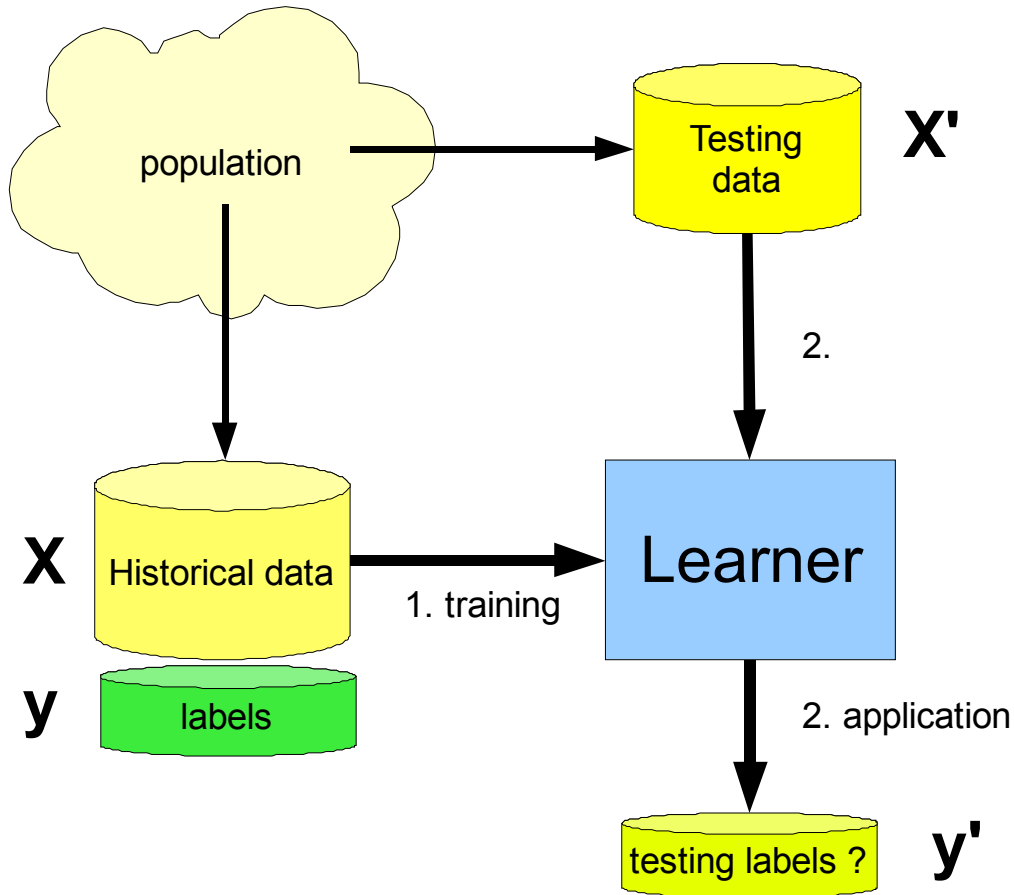
- Adapting with limited resources
 - time and memory



Setting: adaptive learning



Supervised Learning



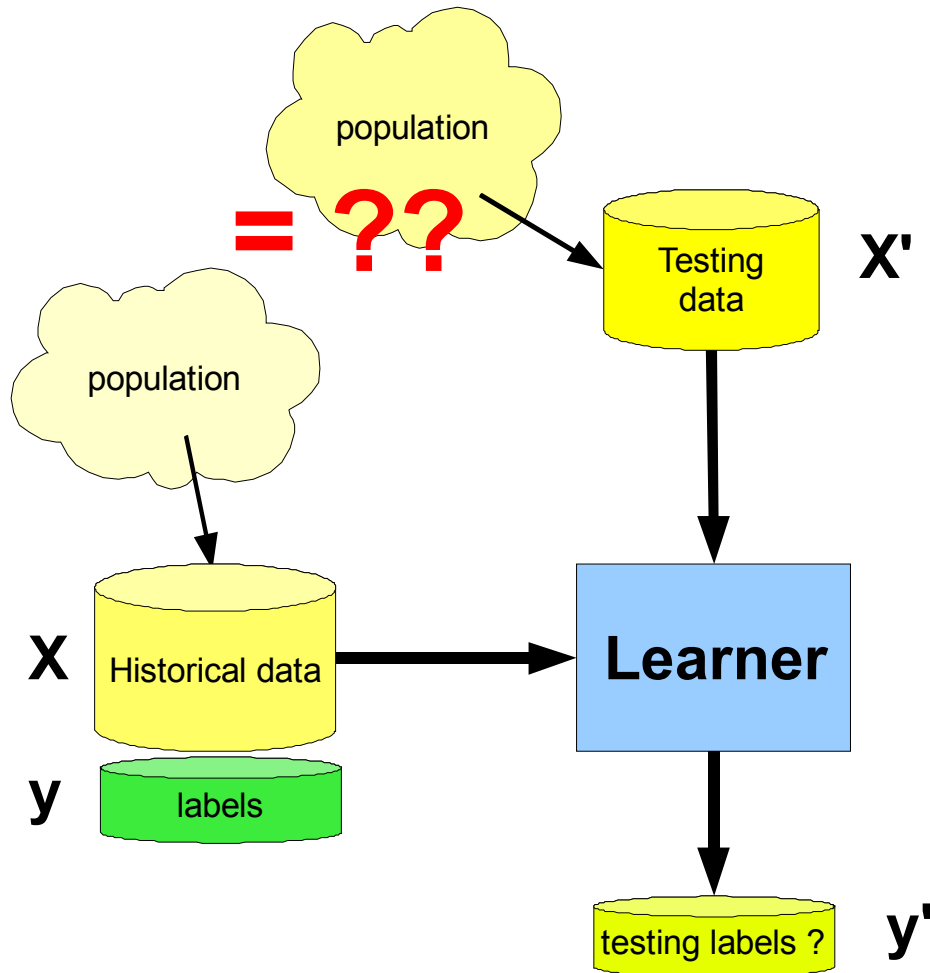
Training:
Learning
a mapping function

$$y = L(x)$$

Application:
Applying L
to unseen data

$$y' = L(X')$$

Learning with concept drift



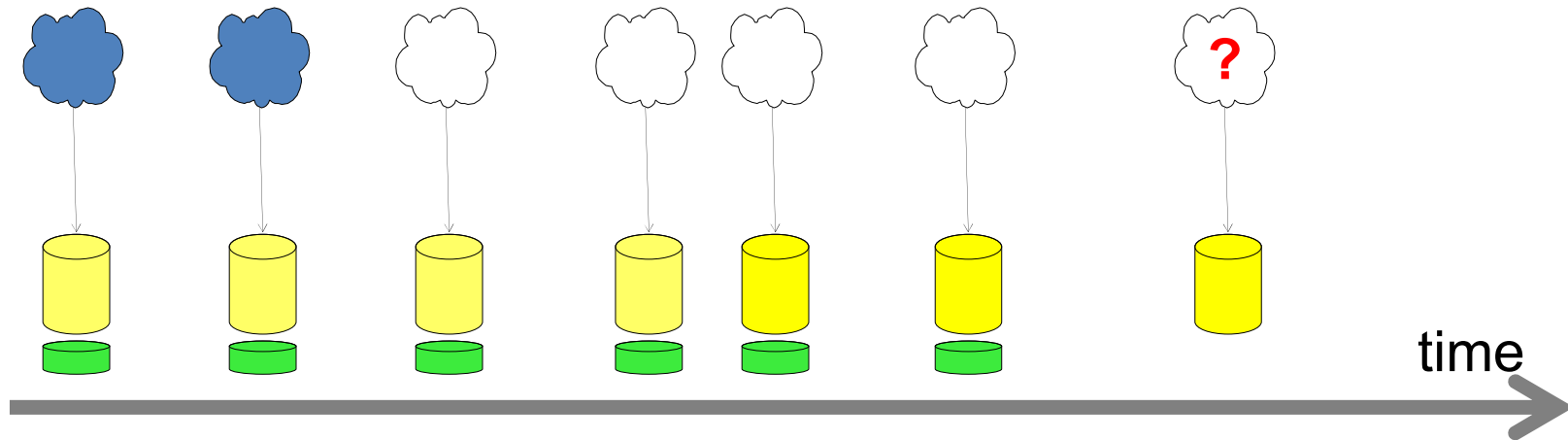
Training:

$$y = L(x)$$

Application:

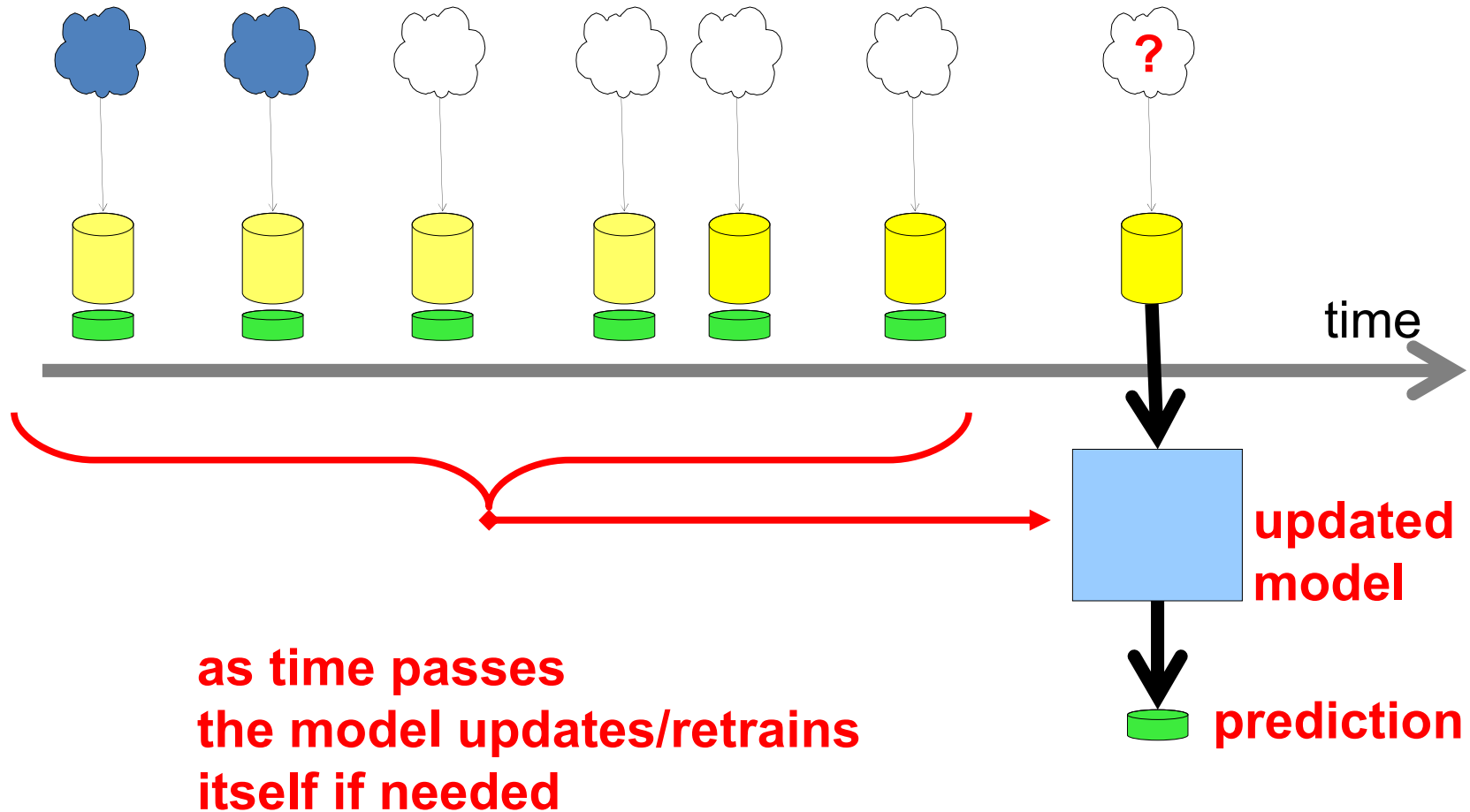
$$y' = L??(X')$$

Online setting

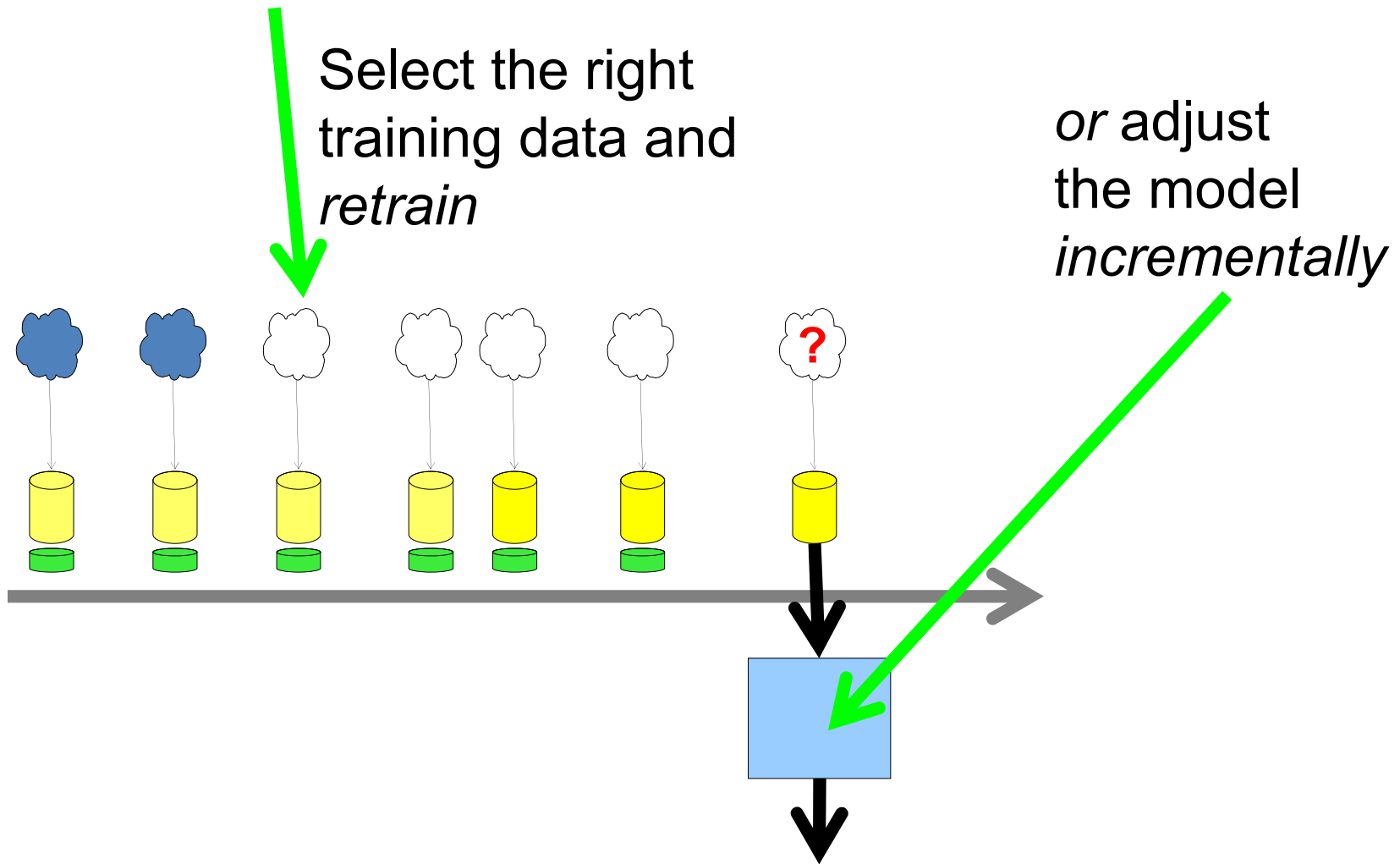


Data arrives in a stream

Adaptive Learning



How to Adapt

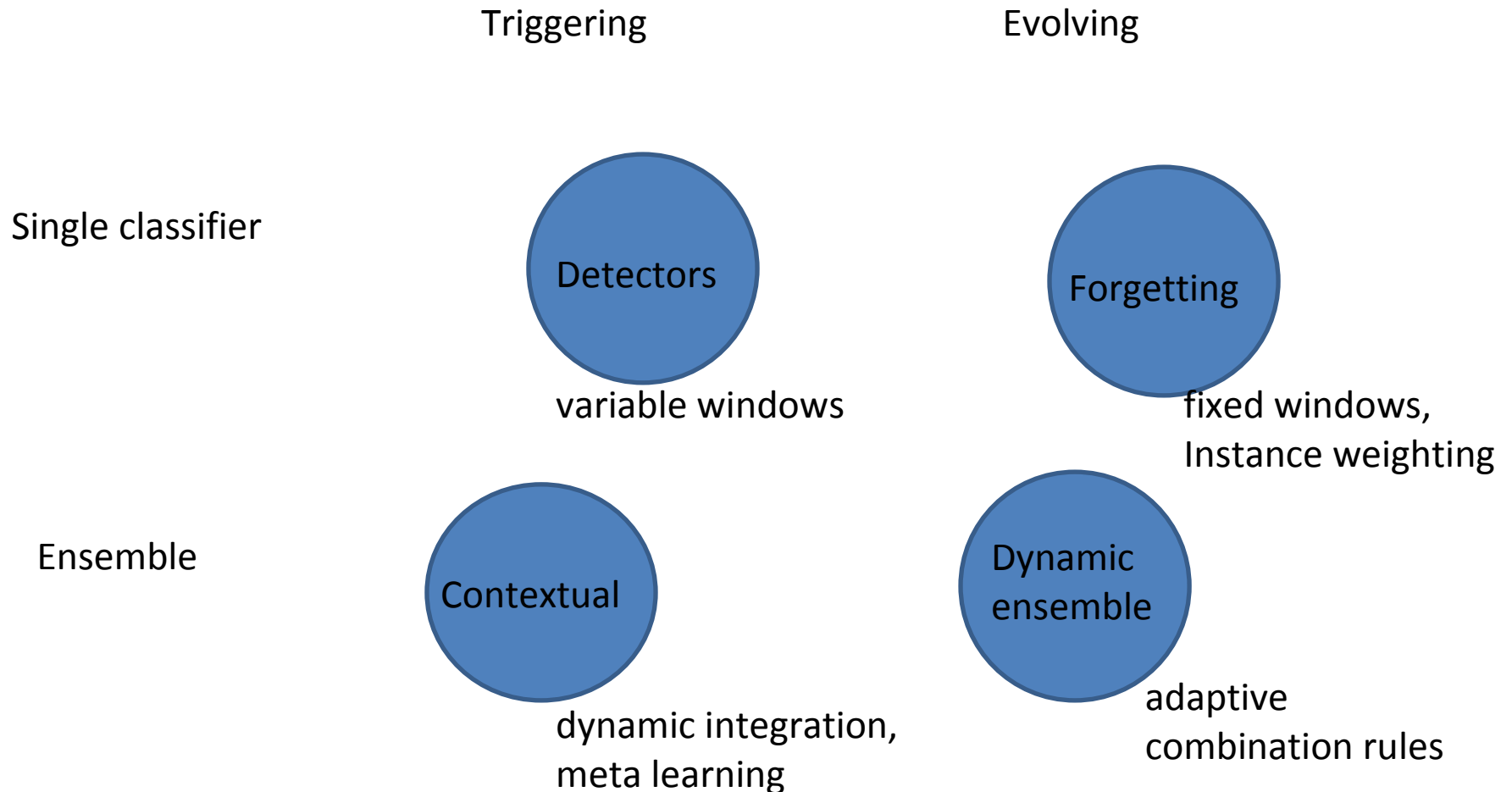


The main strategies how to handle concept drift

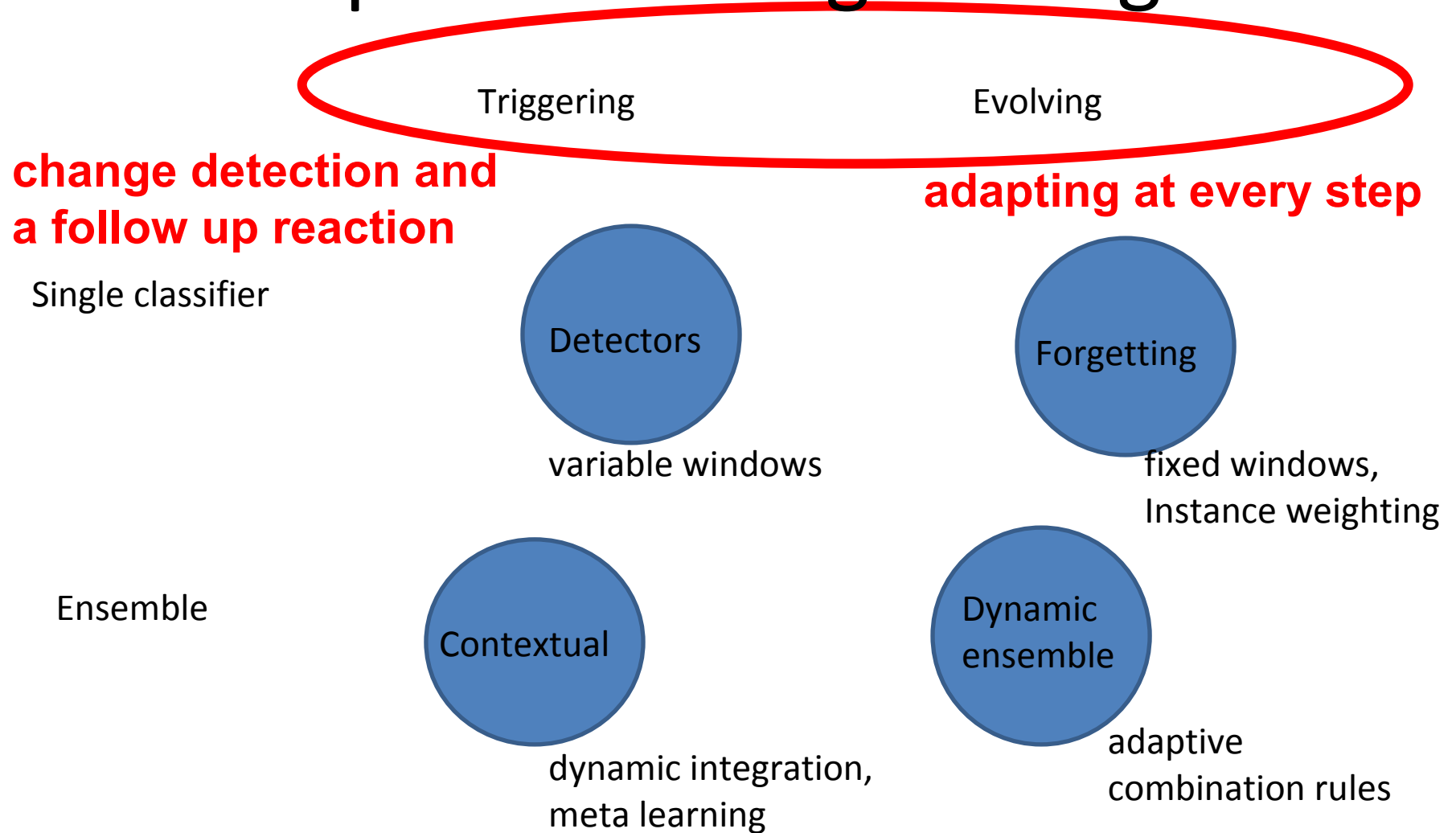


Images.com

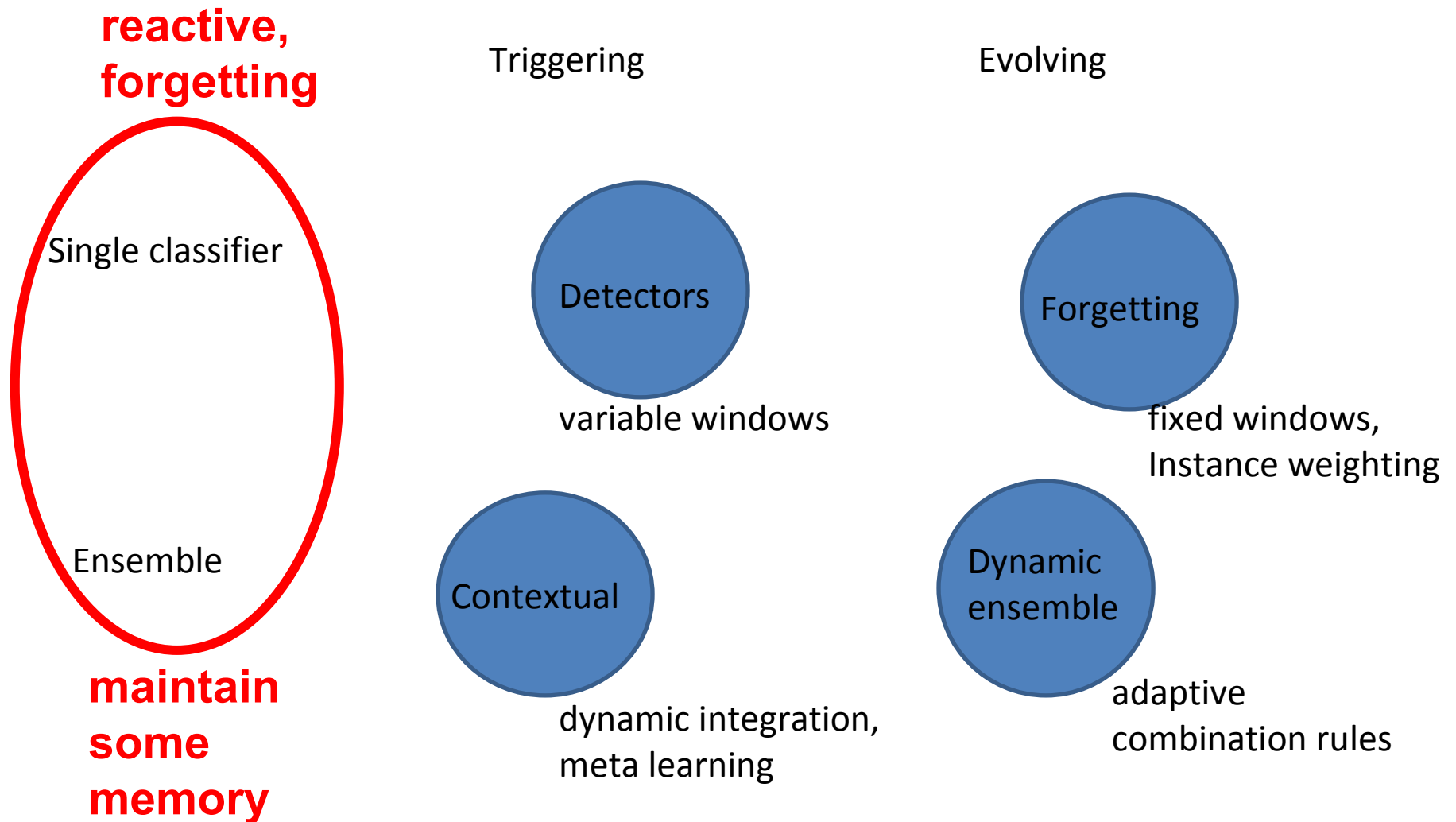
Adaptive learning strategies



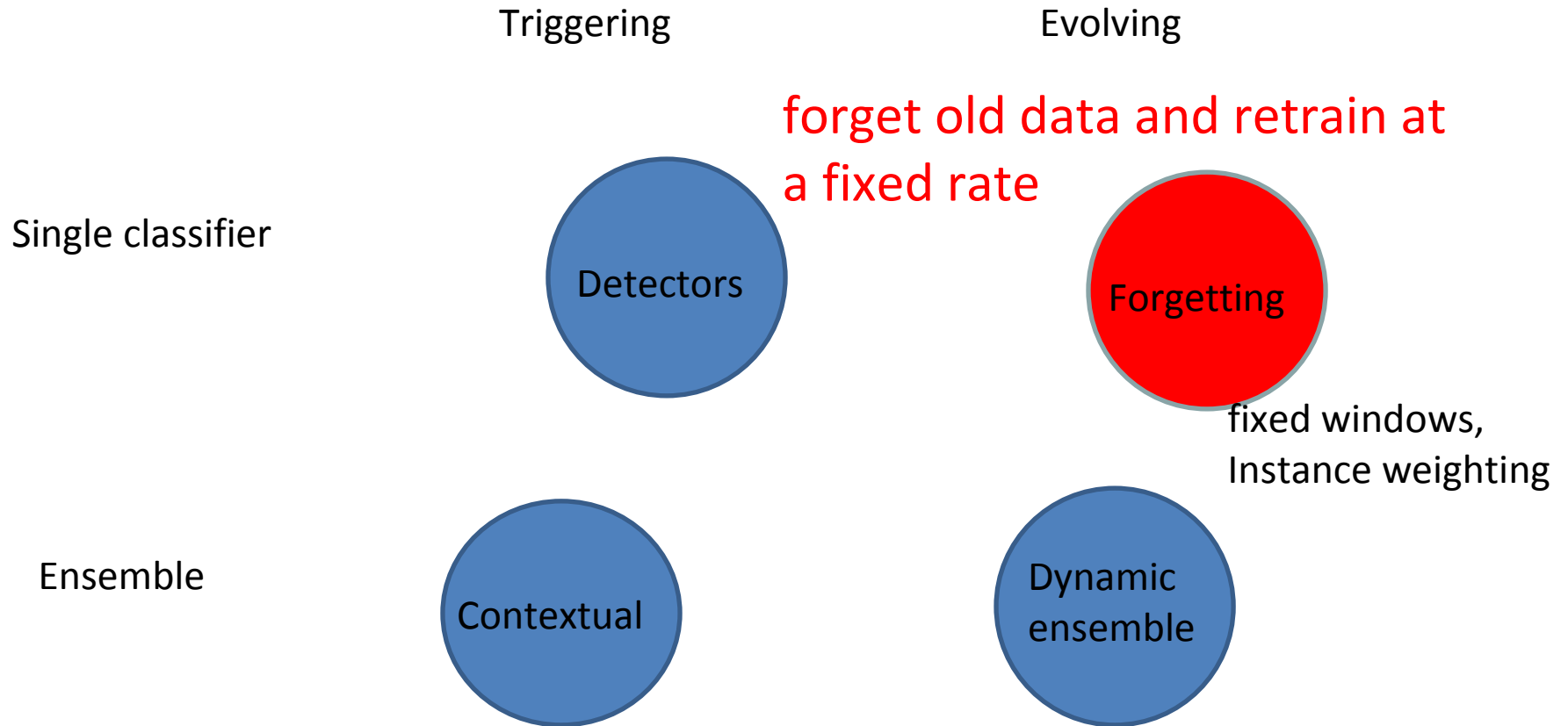
Adaptive learning strategies



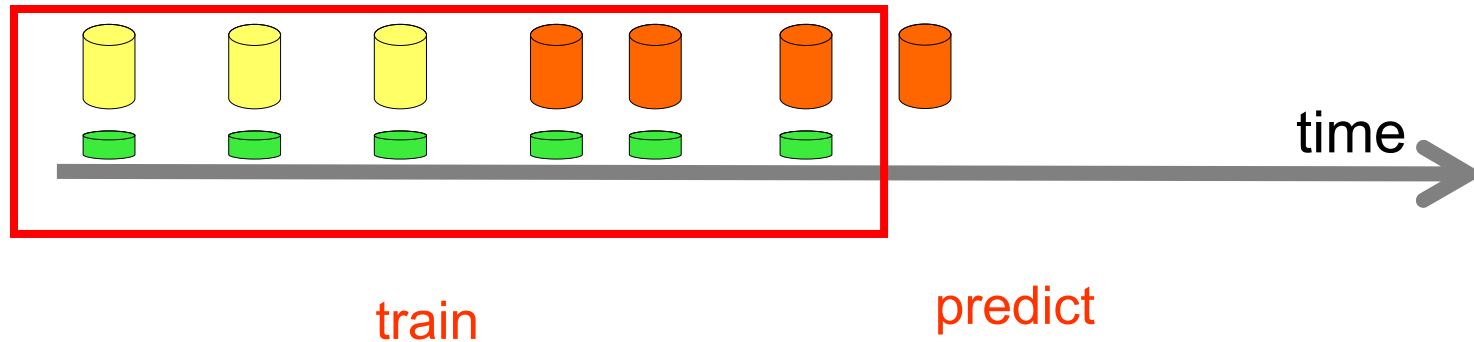
Adaptive learning strategies



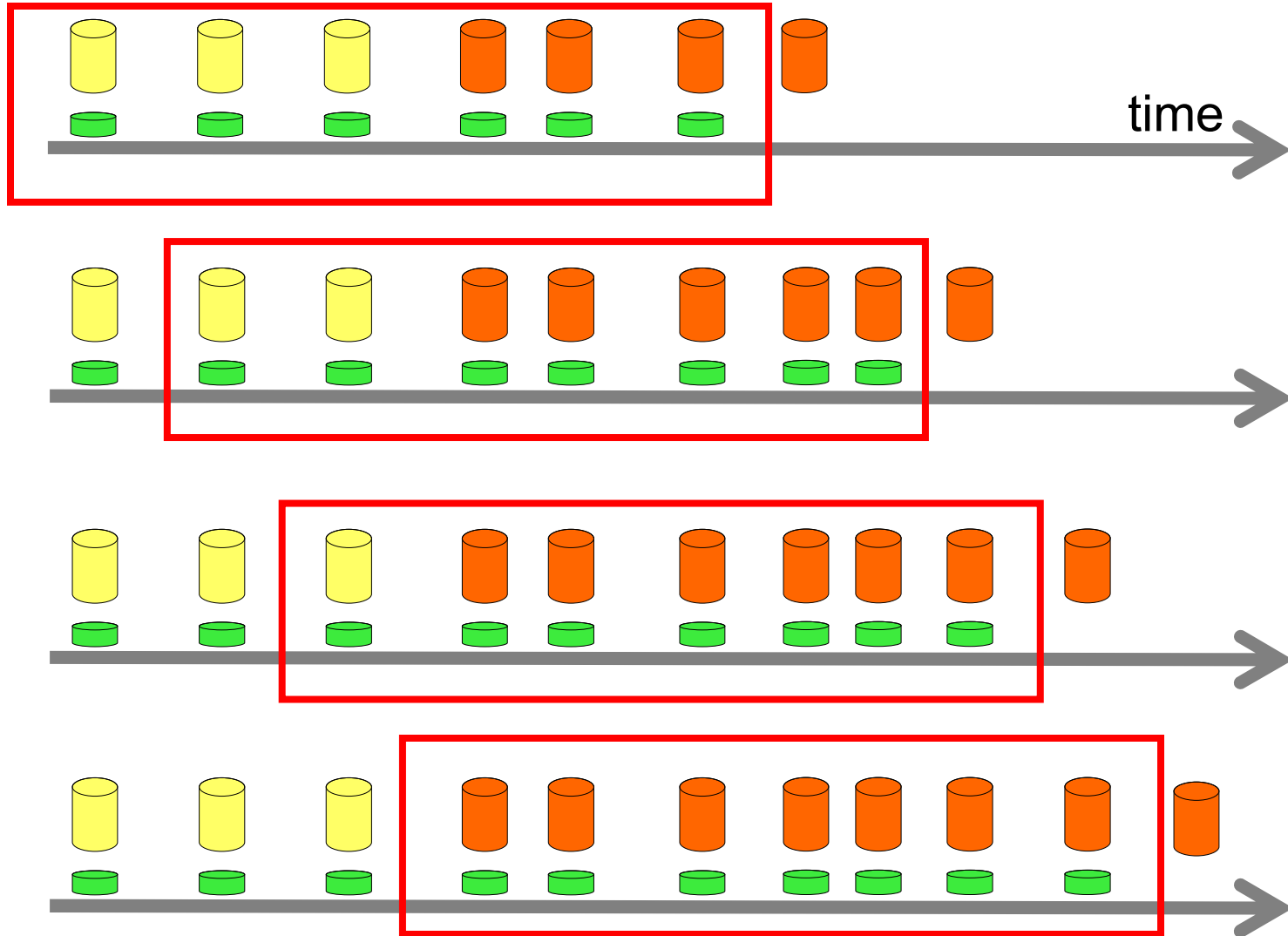
Adaptive learning strategies



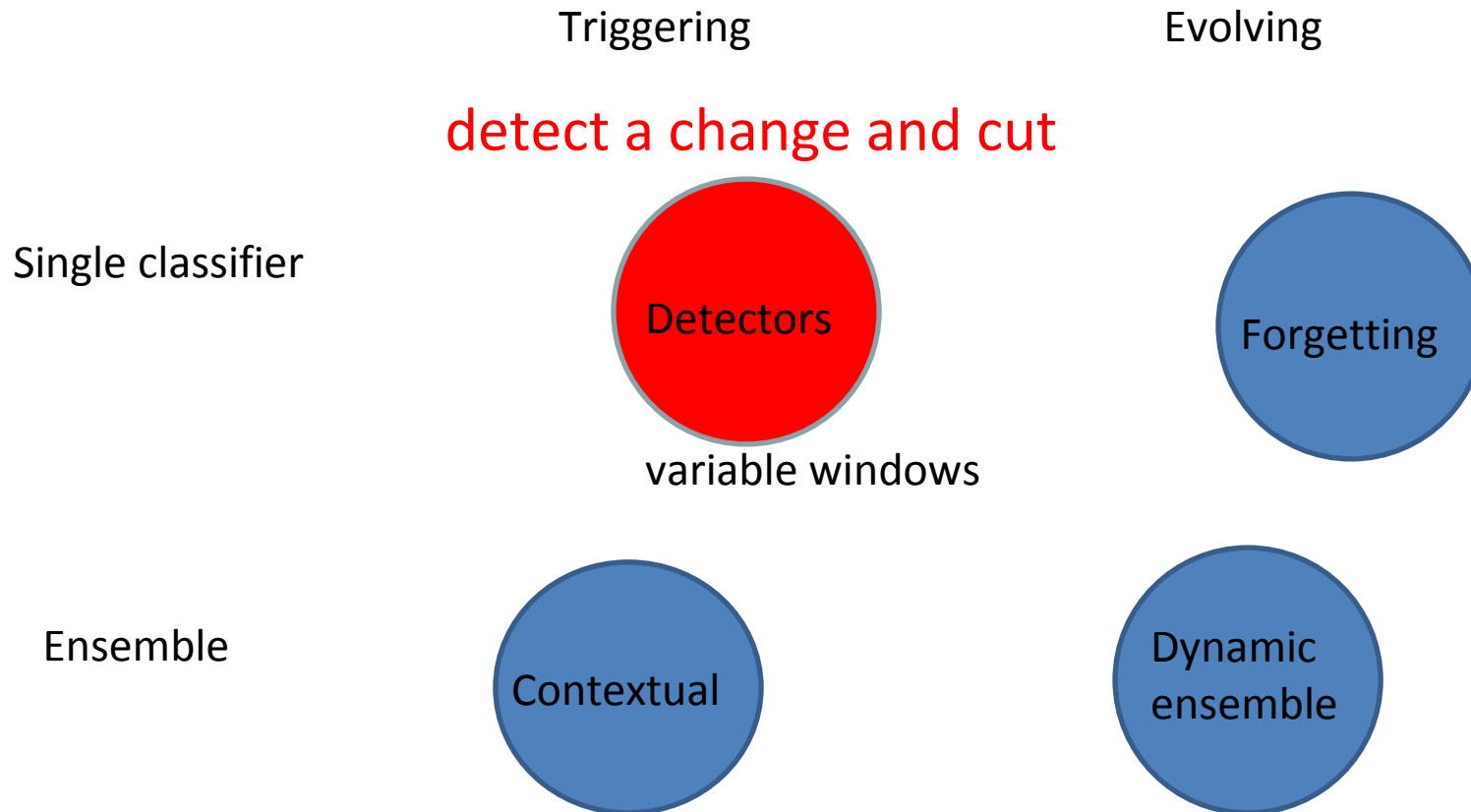
Fixed Training Window



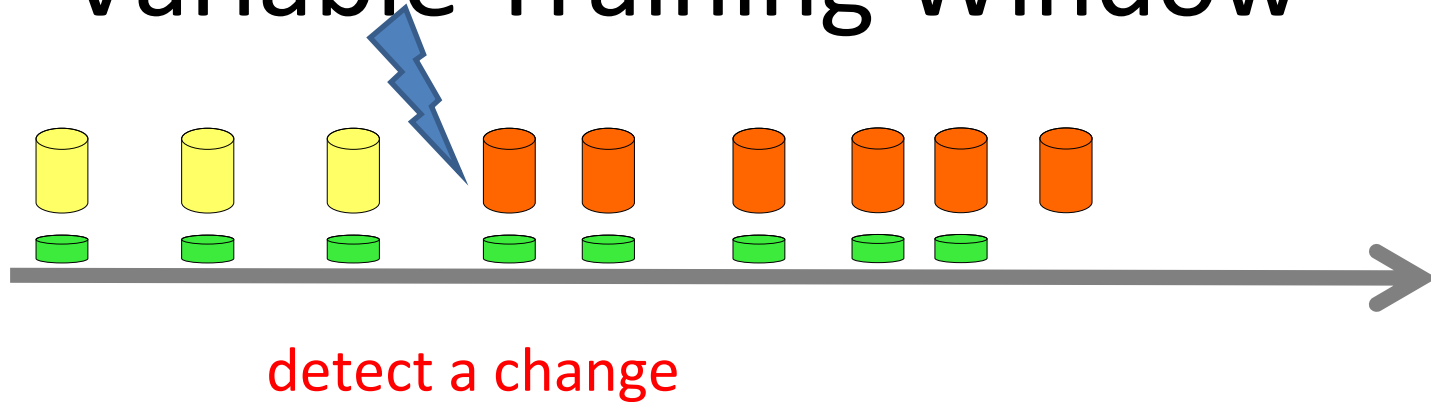
Fixed Training Window



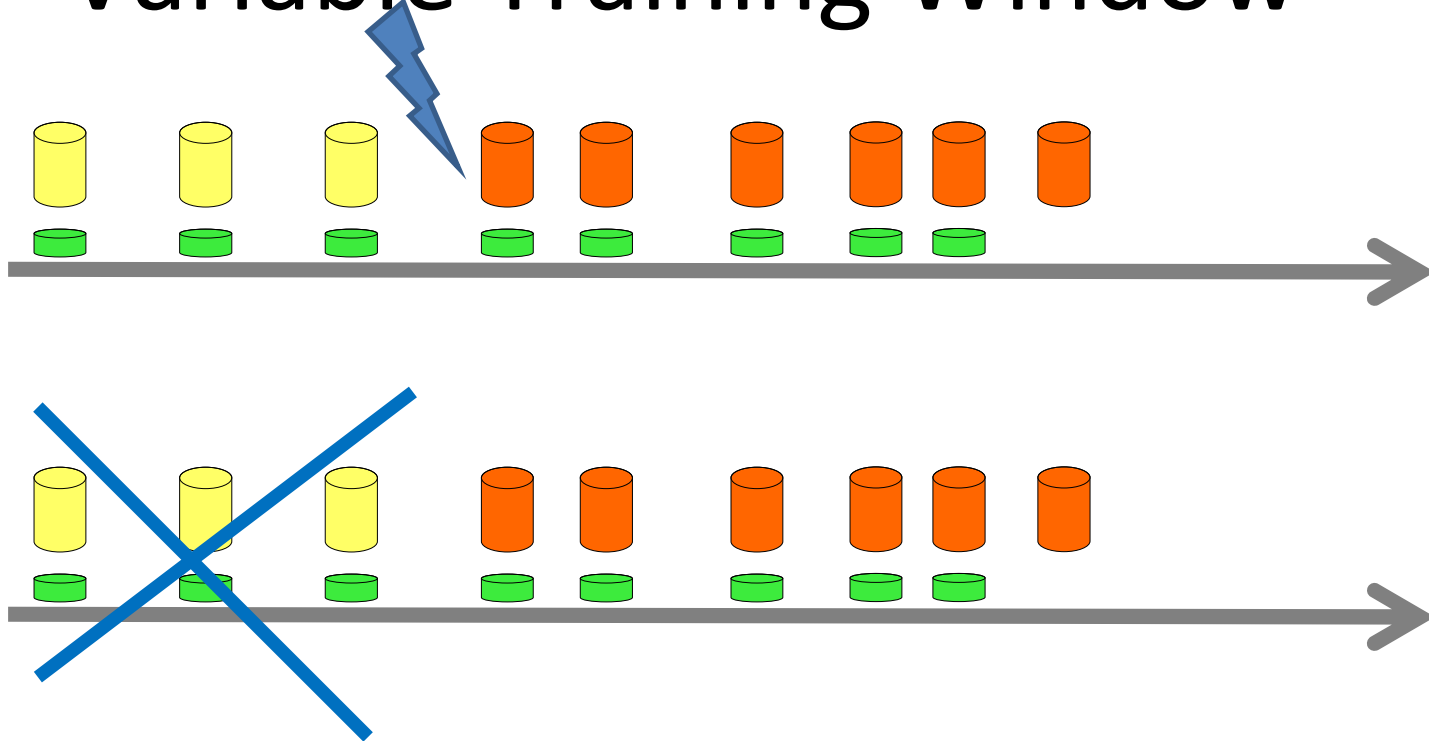
Adaptive learning strategies



Variable Training Window

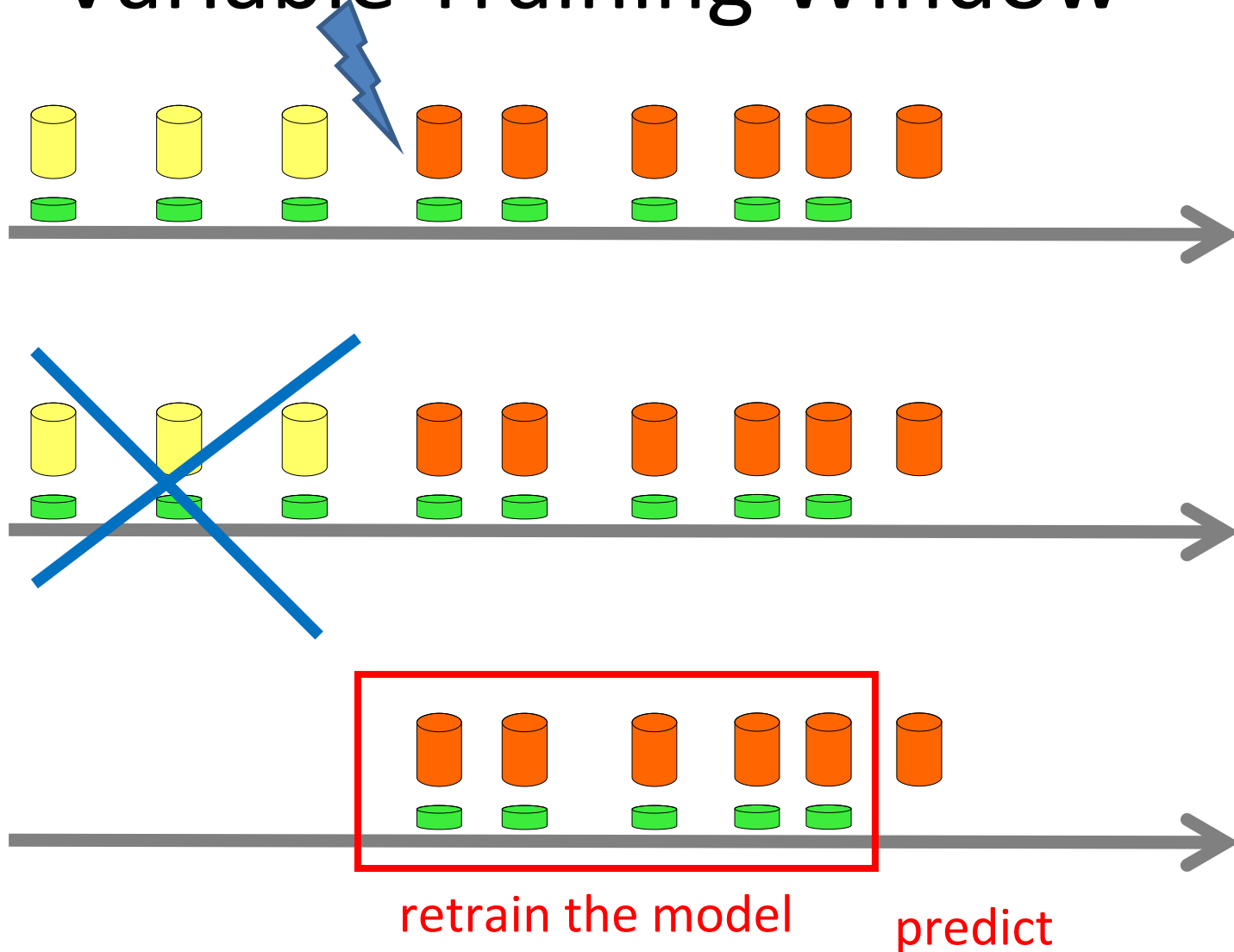


Variable Training Window

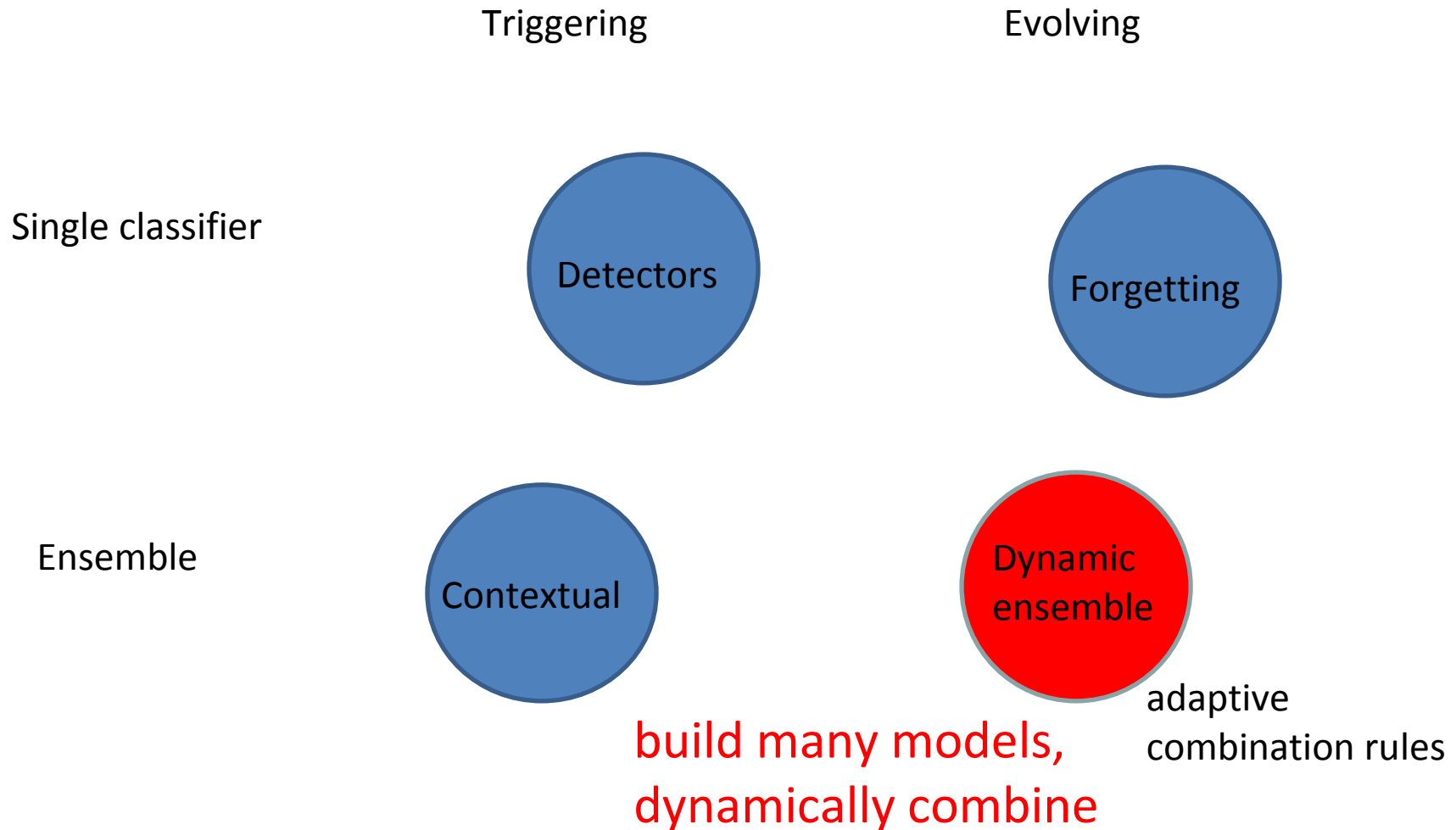


drop old data

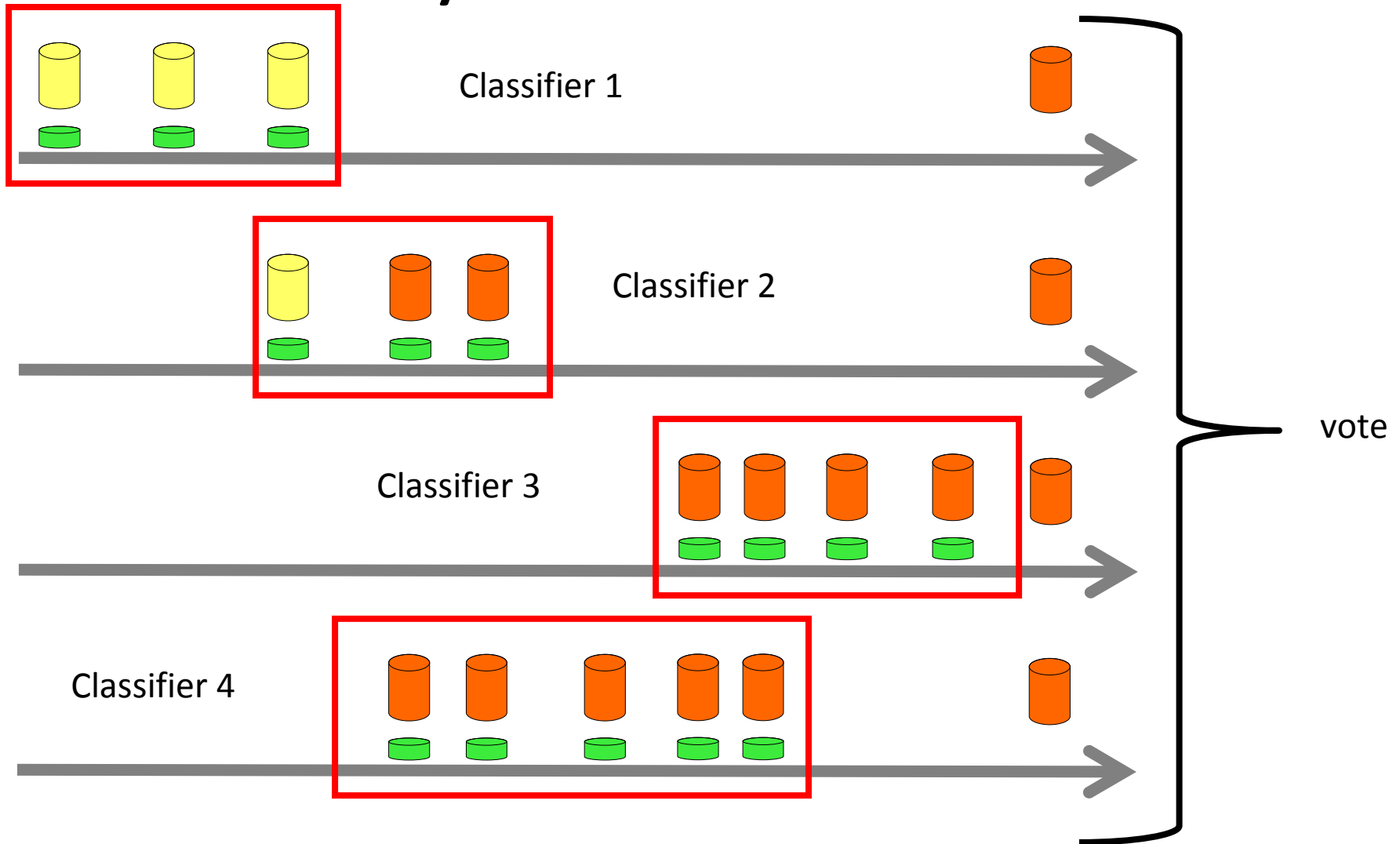
Variable Training Window



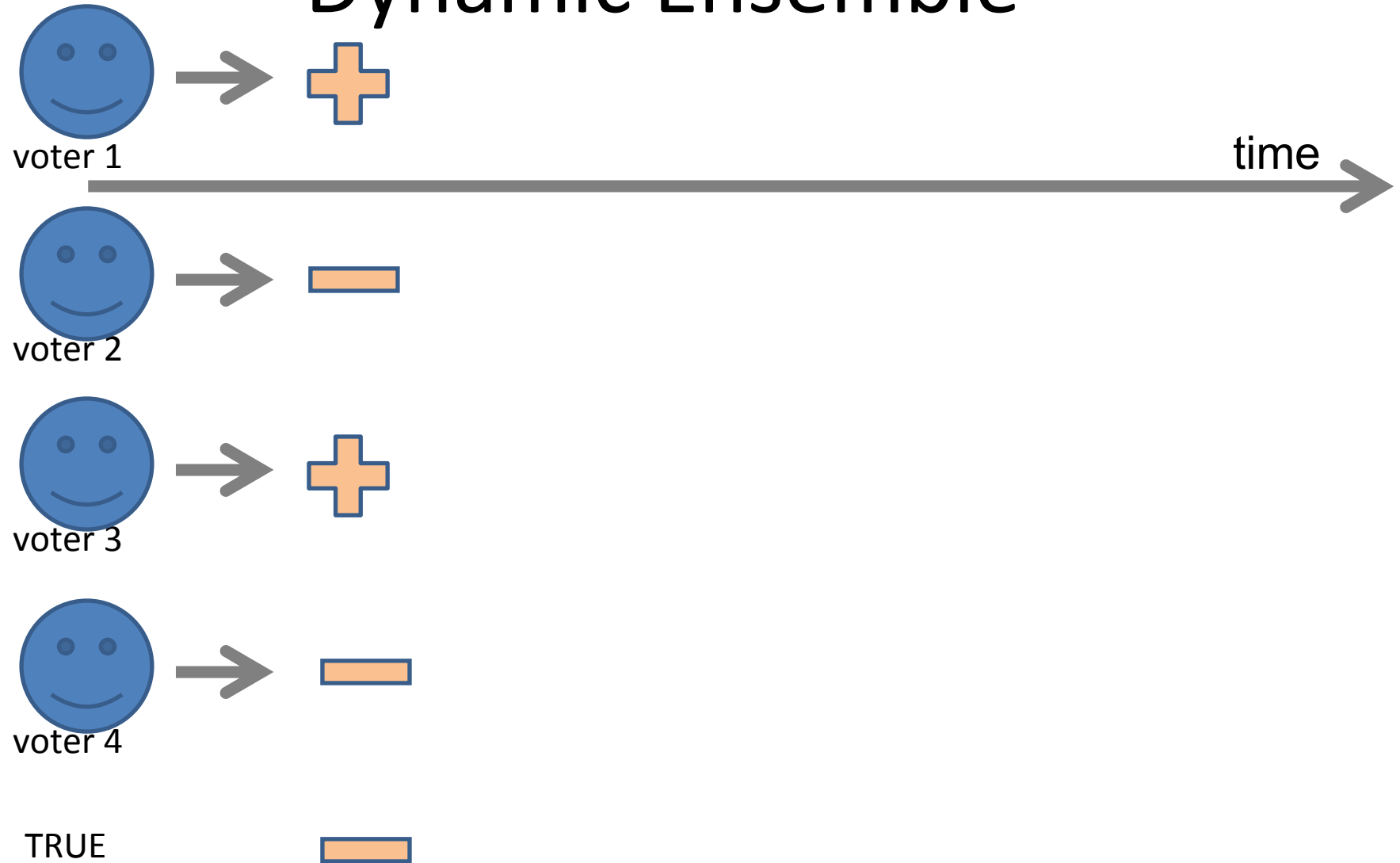
Adaptive learning strategies



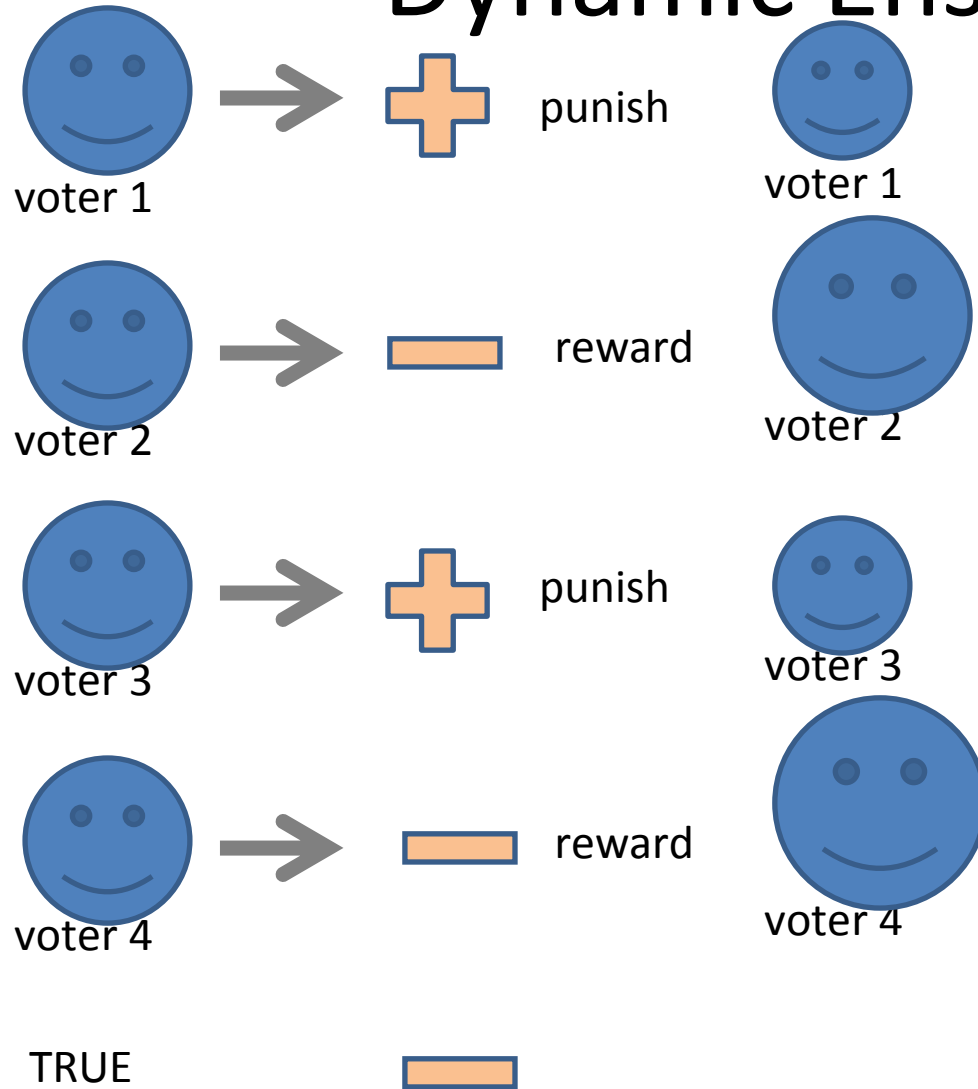
Dynamic Ensemble



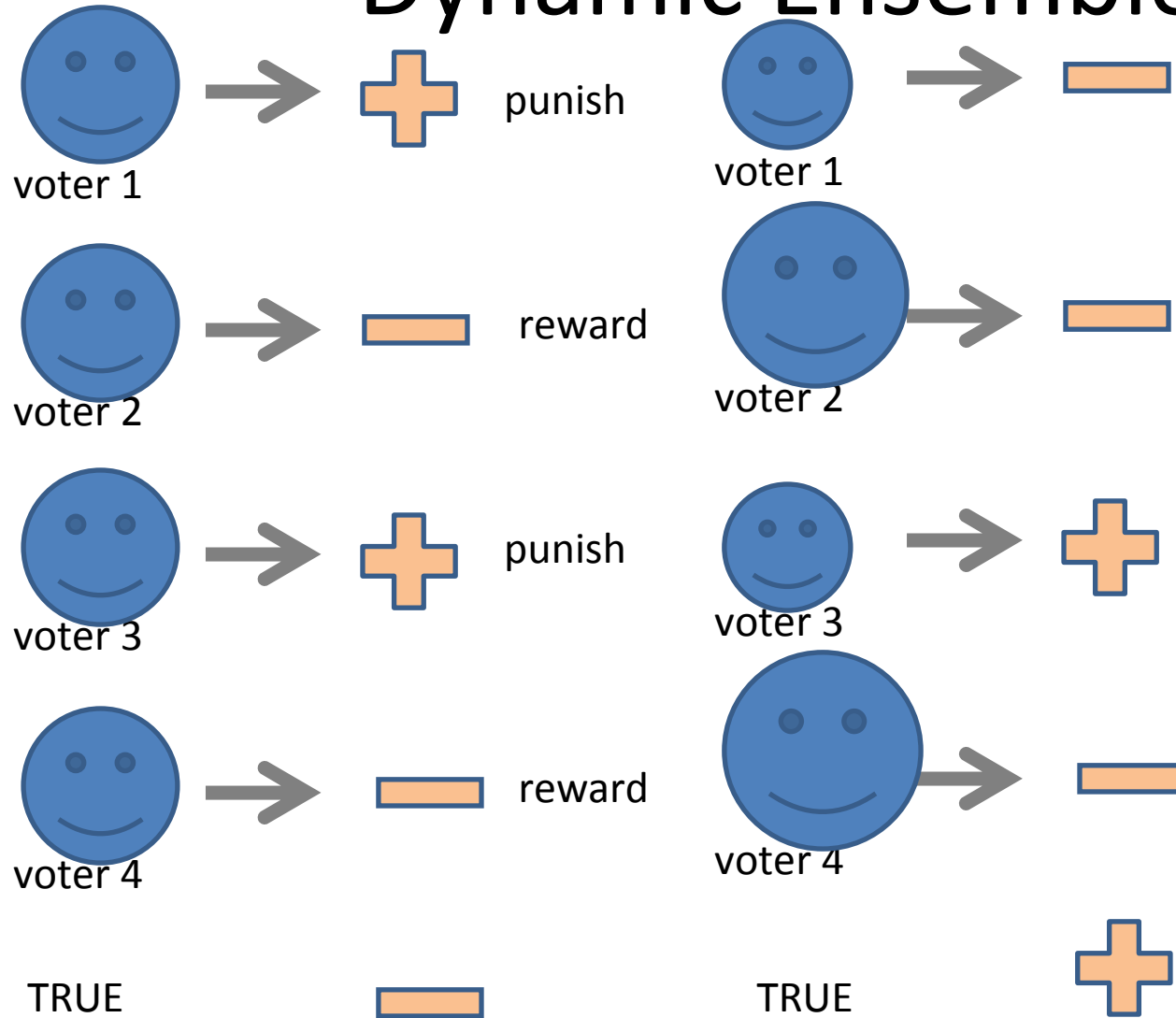
Dynamic Ensemble



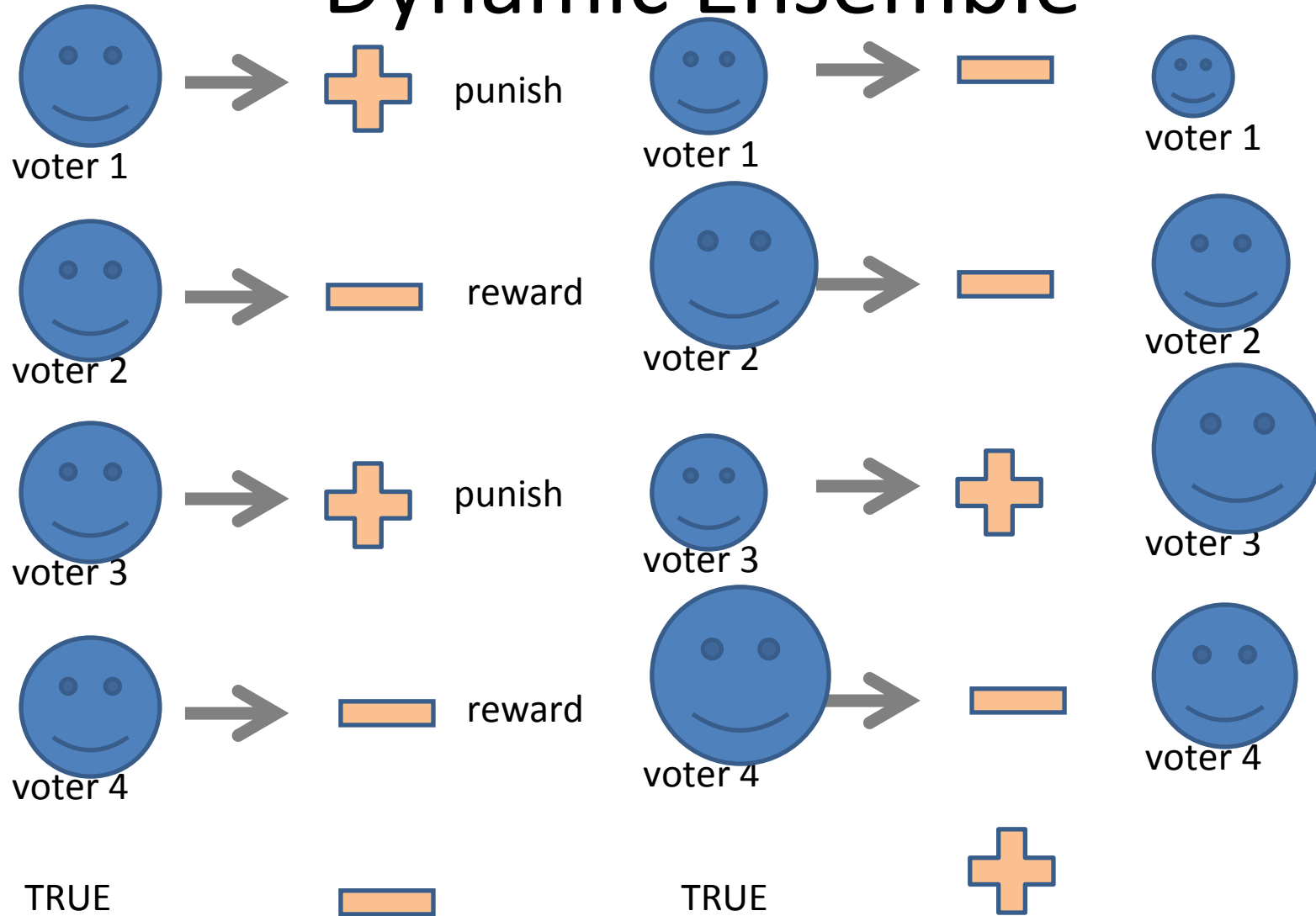
Dynamic Ensemble



Dynamic Ensemble



Dynamic Ensemble

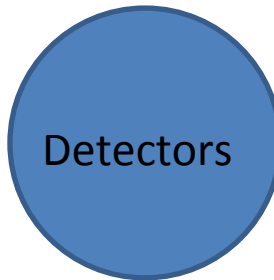


Adaptive learning strategies

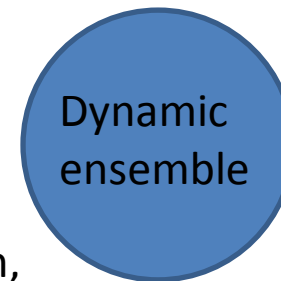
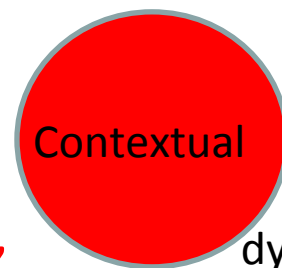
Triggering

Evolving

Single classifier



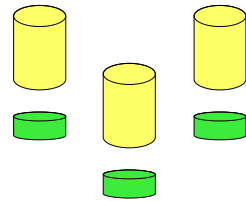
Ensemble



build many models,
switch models according
to the observed incoming data

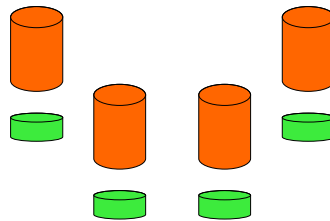
dynamic integration,
meta learning

Contextual (Meta) Approaches



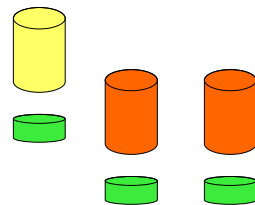
Group 1 = Classifier 1

partition the training data
build classifiers

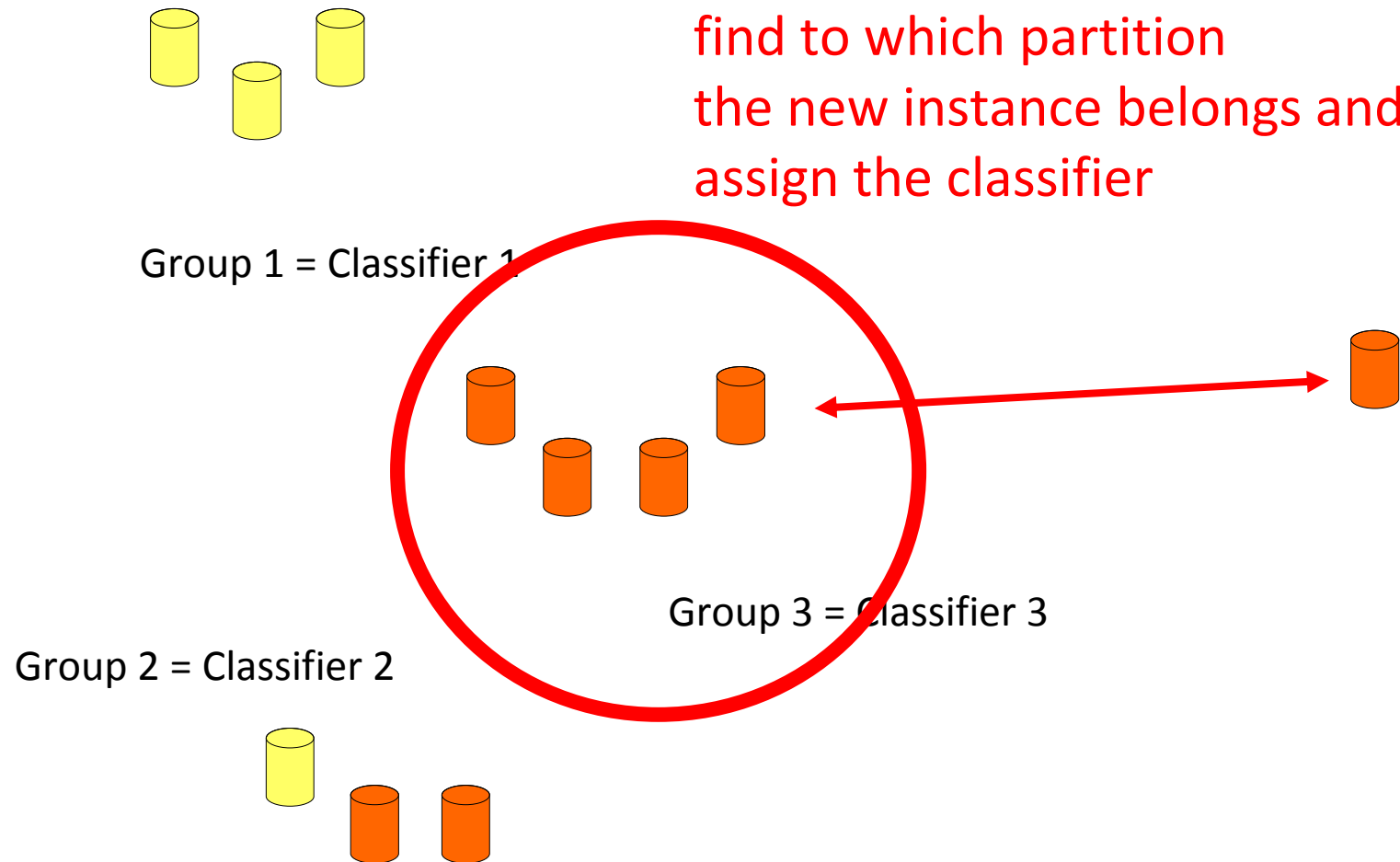


Group 3 = Classifier 3

Group 2 = Classifier 2

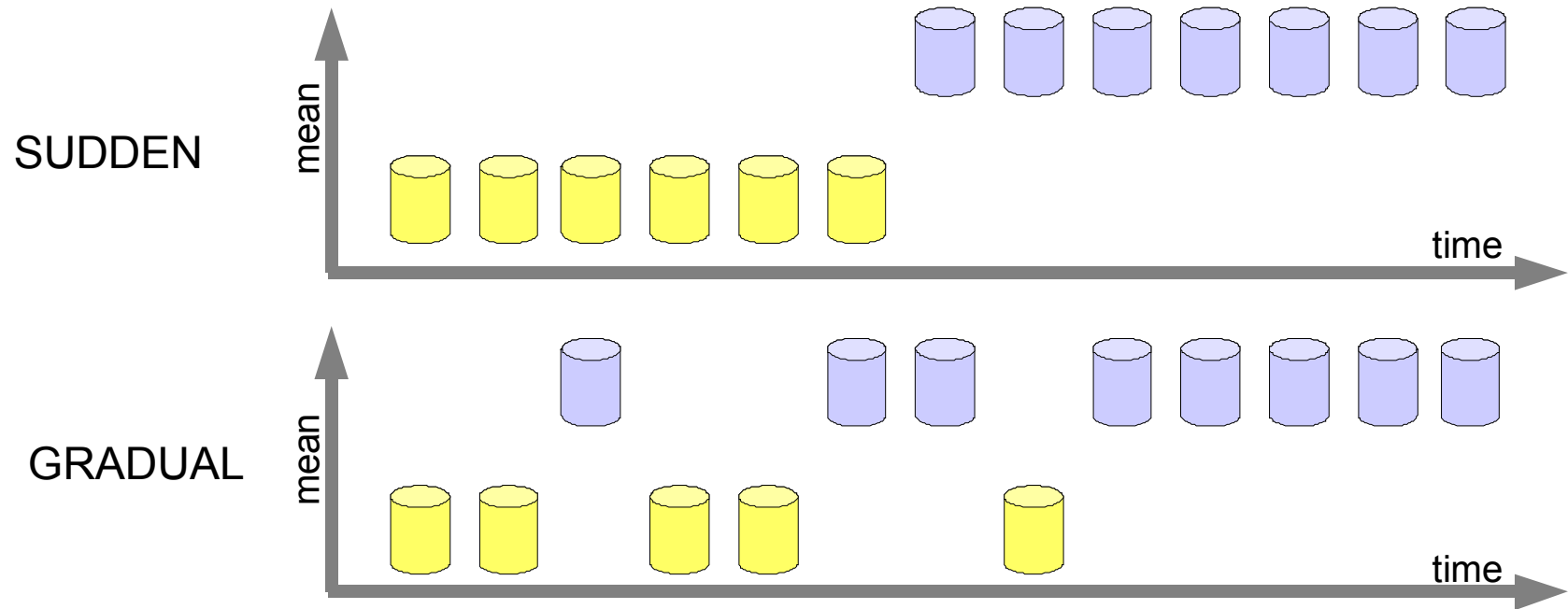


Contextual (Meta) Approaches



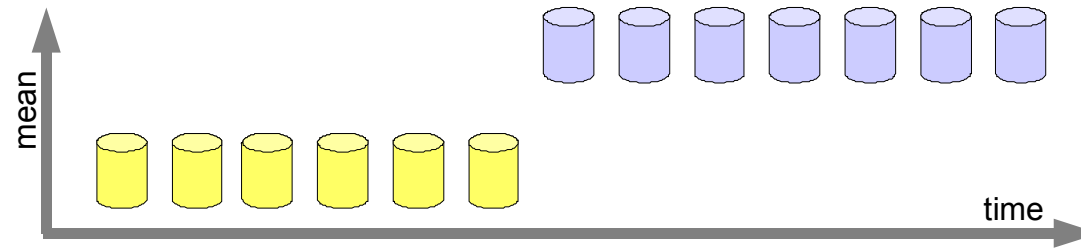
- adaptive learning approaches implicitly or explicitly assume some type of change

Types of Changes: speed

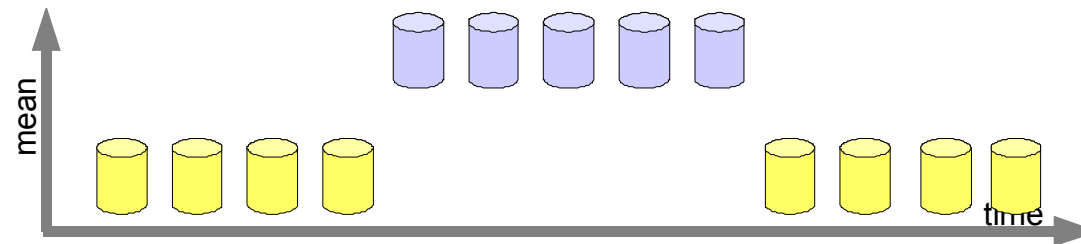


Types of Changes: reoccurrence

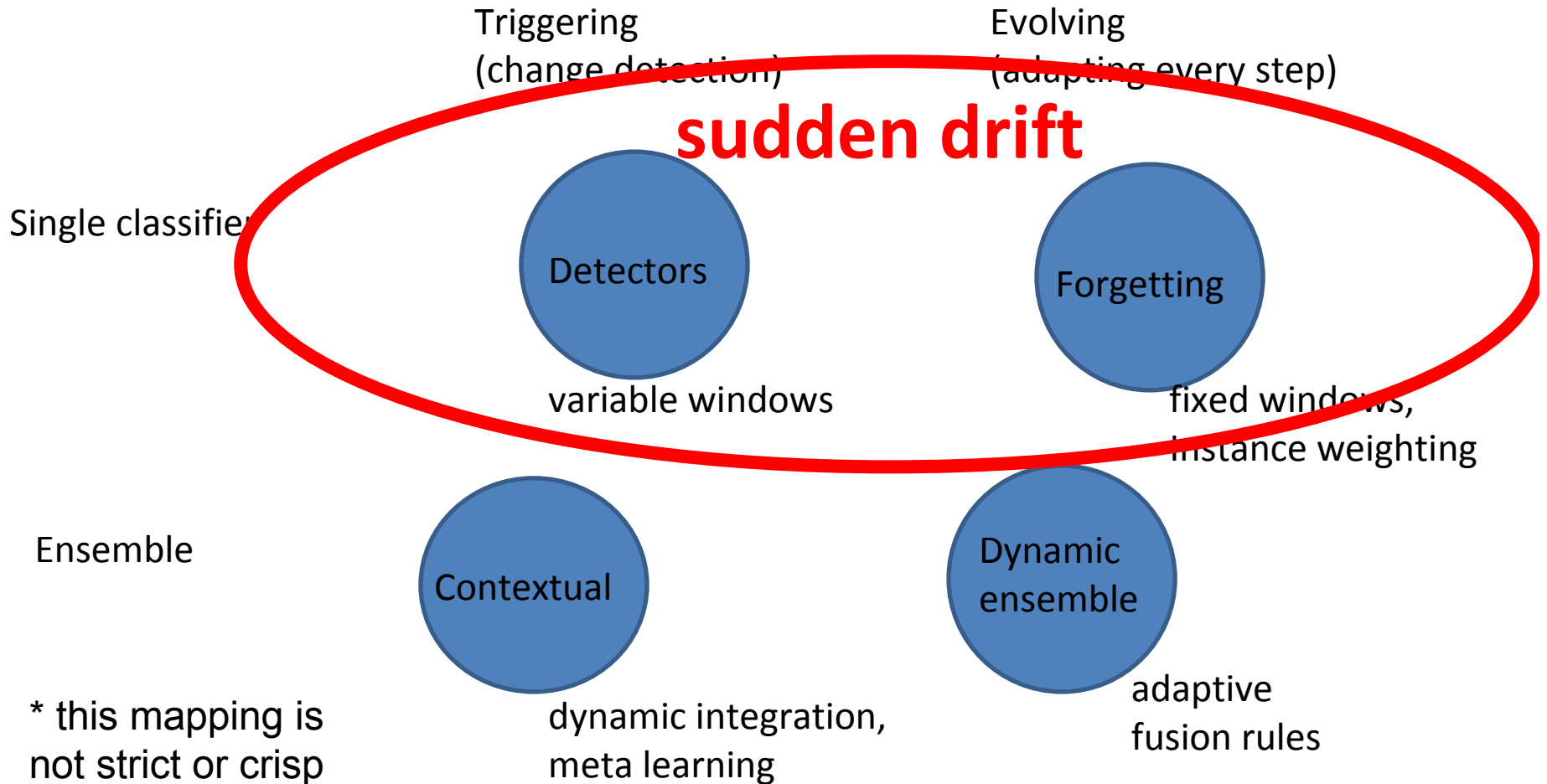
ALWAYS NEW*



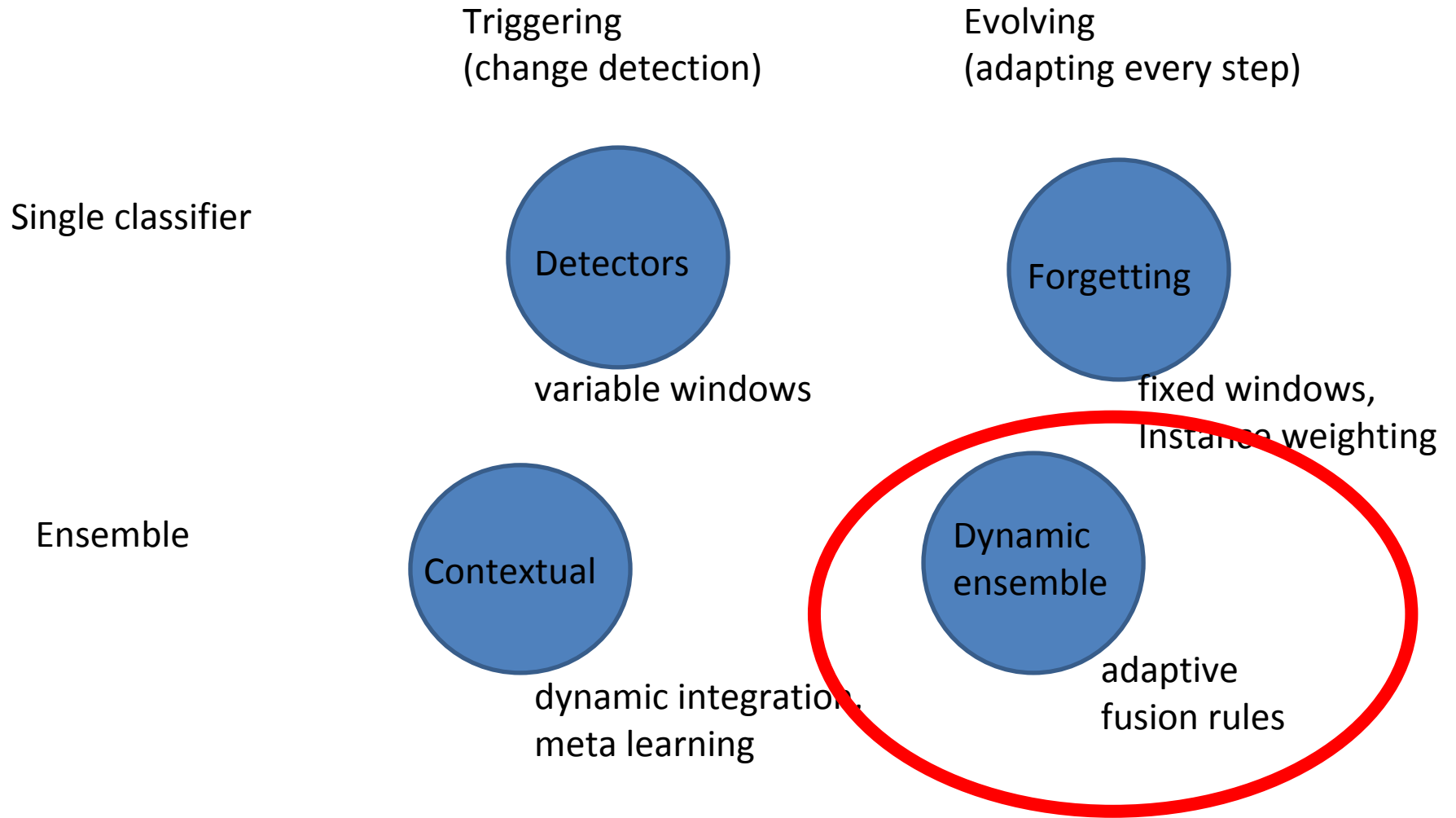
REOCCURING



Typically Used*

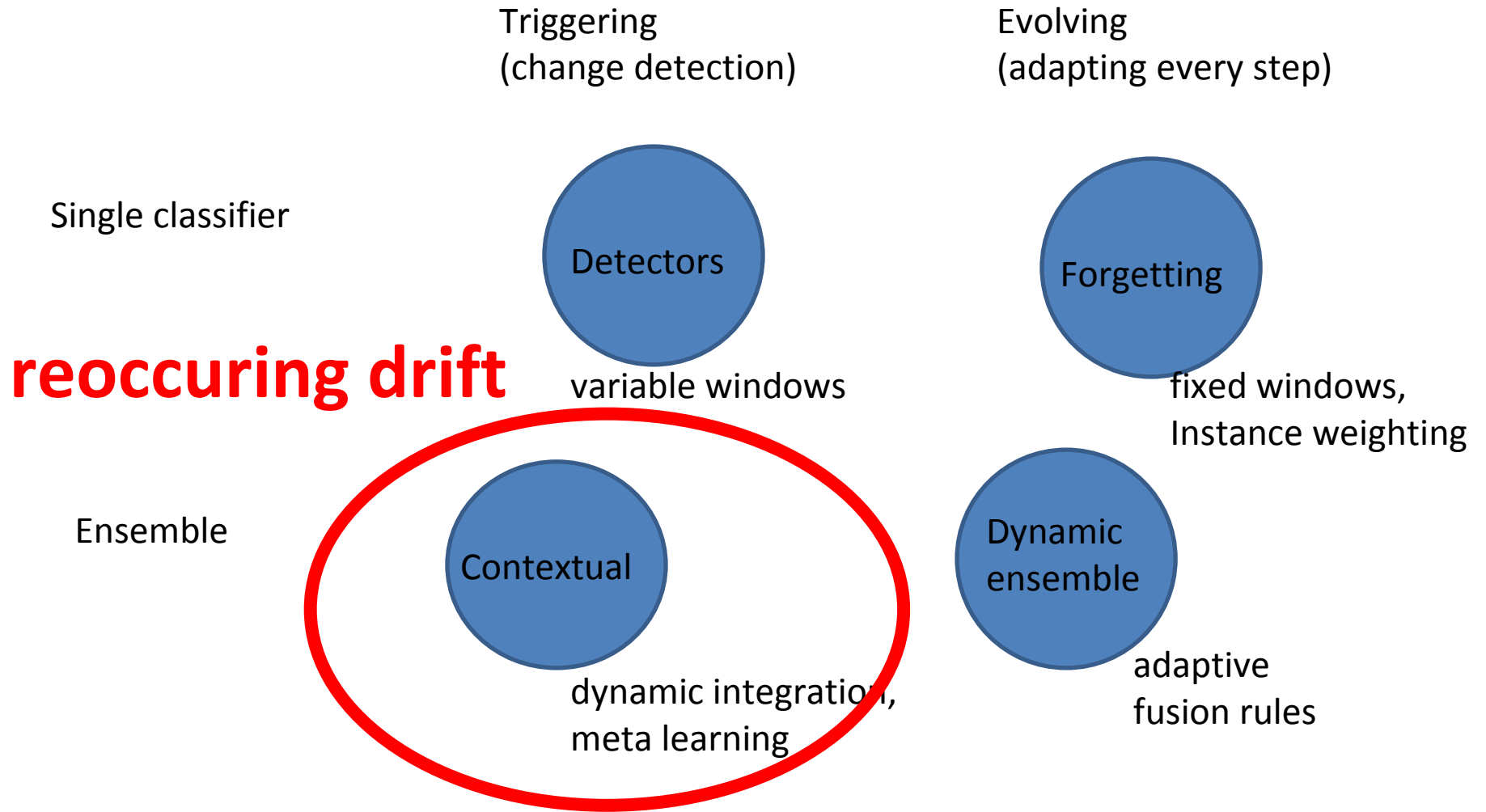


Typically Used



gradual drift

Typically Used



Which approach to use?

- changes occur over time
- we need models that evolve over time
- choice of technique depends on
 - what type of change is expected
 - user goals/ applications

Block summary



Images.com

Block summary

- Concept drift is a problem in data stream applications, since static models lose accuracy
- Different adaptive techniques exist
- The choice of technique depends on
 - what change is expected
 - peculiarities of the task and application

Block 2: methods and techniques

The goal:

to discuss selected popular approaches from each of the major types in more detail

- Data management

Forgetting

- Detection Methods

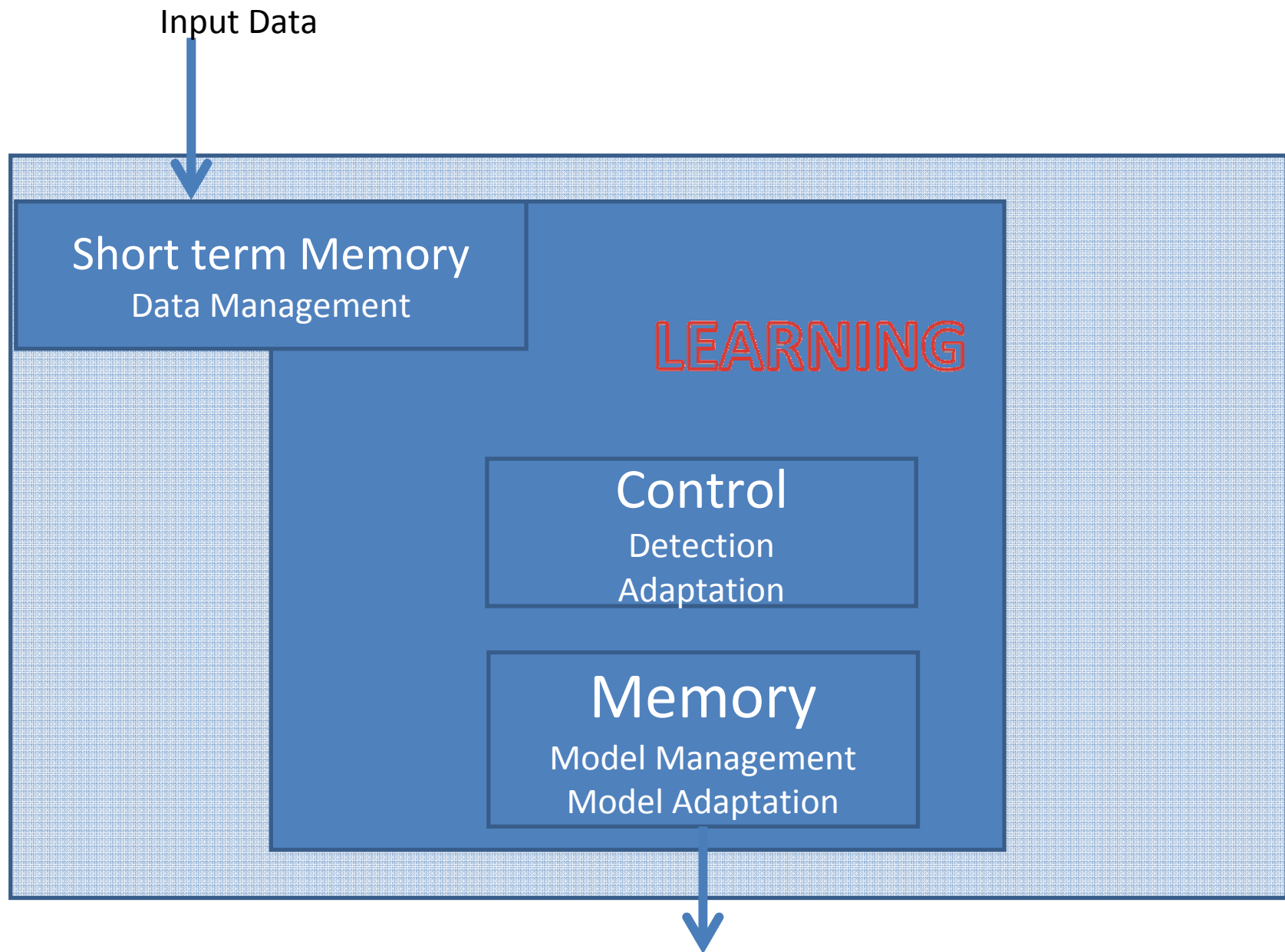
Detectors

- Adaptation methods

Dynamic ensembles

- Decision model management

Contextual



Data Management

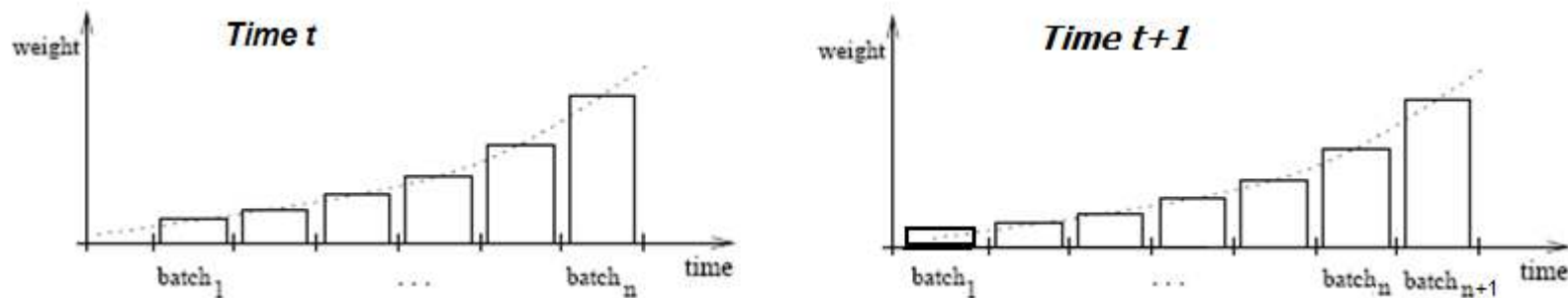
- Characterize the information stored in memory to maintain a decision model consistent with the actual state of the nature.
 - Full Memory.
 - Partial Memory.

Weighting examples

- Full Memory.
 - Store in memory sufficient statistics over all the examples.
 - Weighting the examples accordingly to their age.
 - Oldest examples are less important.
- Weighting examples based on the age:

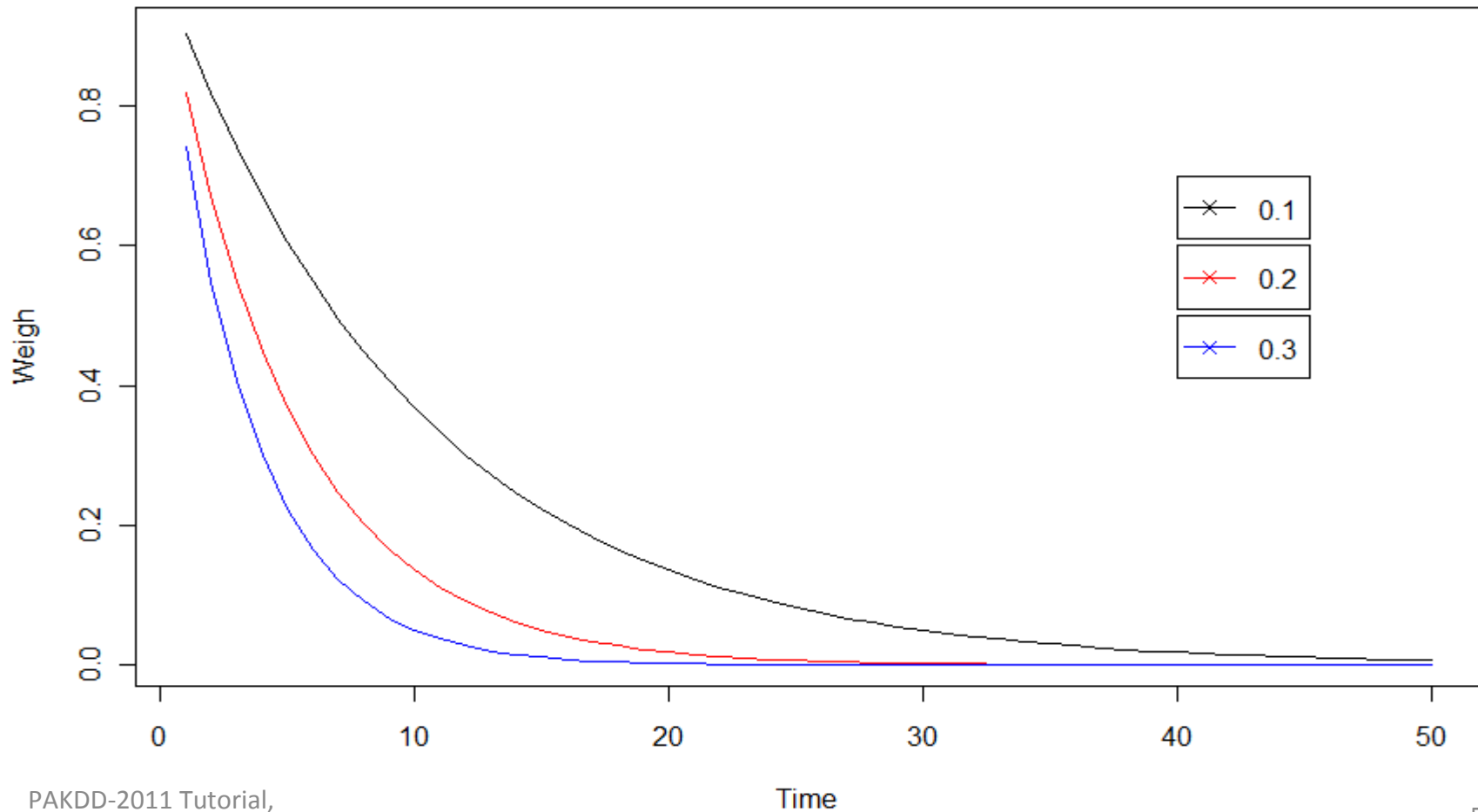
$$w_{\lambda}(x) = \exp(-\lambda t_x)$$

where example \mathbf{x} was found t_x time steps ago



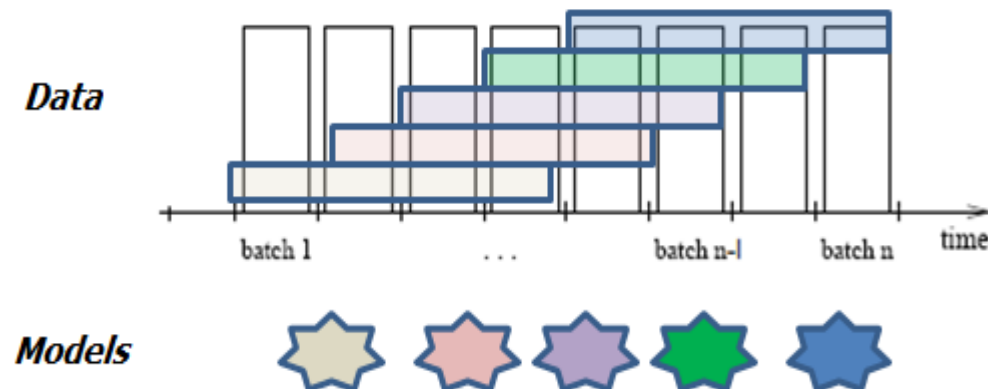
Weight as a Function of Age

Weighting Examples



Partial Memory

- Partial Memory.
 - Store in memory only the most recent examples.
 - Examples are stored in a *first-in first-out* data structure.
 - At each time step the learner induces a decision model using only the examples that are included in the window.



Partial Memory

- The key difficulty is how to select the appropriate window size:
 - Small window
 - can assure a fast adaptability in phases with concept changes
 - In more stable phases it can affect the learner performance
 - Large window
 - produce good and stable learning results in stable phases
 - can not react quickly to concept changes.

Partial Memory - Windows

- Fixed Size windows.
 - Store in memory a fixed number of the most recent examples.
 - Whenever a new example is available:
 - it is stored in memory and
 - the oldest one is discarded.
 - This is the simplest method to deal with concept drift and can be used as a baseline for comparisons.
- Adaptive Size windows.
 - the set of examples in the window is variable.
 - They are used in conjunction with a detection model.
 - Decreasing the size of the window whenever the detection model signals drift and increasing otherwise.

Discussion

- The data management model also indicates the forgetting mechanism.
 - Weighting examples:
 - Corresponds to a gradual forgetting.
 - The relevance of old information is less and less important.
 - Time windows
 - Corresponds to abrupt forgetting;
 - The examples are deleted from memory.
 - Of course we can combine both forgetting mechanisms by weighting the examples in a time window.

Discussion

- These methods are *blind* adaptation models:
 - There is no explicit change detection
 - Do not provide:
 - Any indication about change points
 - Dynamics of the process generating data

Detection Methods

- The Detection Model characterizes the techniques and mechanisms for explicit drift detection.
 - An advantage of the detection model is they can provide:
 - Meaningful descriptions:
 - indicating change-points or
 - small time-windows where the change occurs
 - Quantification of the changes.

Detection Methods

- Monitoring the evolution of performance indicators.
(See Klinkenberg (98) for a good overview of these indicators)
 - Performance measures,
 - Properties of the data, etc
- Monitoring distributions on two different time-windows
(See D.Kifer, S.Ben-David, J.Gehrke; VLDB 04)
 - A reference window over past examples
 - A window over the most recent examples

Monitoring the evolution of performance indicators

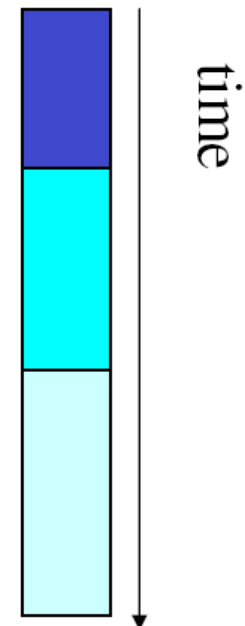
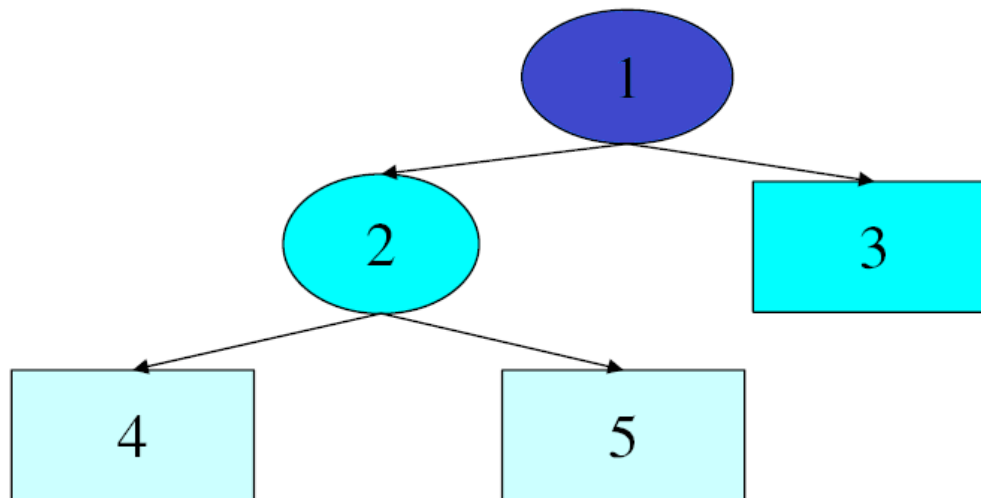
- Klinkenberg (1998) proposed monitoring the values of three performance indicators:
 - Accuracy, recall and precision over time,
 - and then, comparing it to a confidence interval of standard sample errors
 - for a moving average value using the last M batches of each particular indicator.
- FLORA family of algorithms (Widmer and Kubat, 1996) monitors accuracy and model size.
 - includes a window adjustment heuristic for a rule-based classifier.
 - FLORA3: dealing with recurring concepts.

Monitoring distributions on two different time-windows

- D.Kifer, S.Ben-David, J.Gehrke (VLDB 04)
 - Propose algorithms (statistical tests based on Chernoff bound) that examine samples drawn from two probability distributions and decide whether these distributions are different.
- Gama, Fernandes, Rocha: VFDTc (IDA 2006)
 - For streaming data decision tree induction
 - Exploits tree structure: nested trees
 - Continuously monitoring differences between two class-distribution of the examples:
 - the distribution when a node was built and the class-distribution when a node was a leaf and
 - the weighted sum of the class-distributions in the leaves descendant of that node.

Hoeffding Trees

- Multiple windows
 - Each node receives information from different windows
 - Root node: oldest data
 - Leaves: most recent data

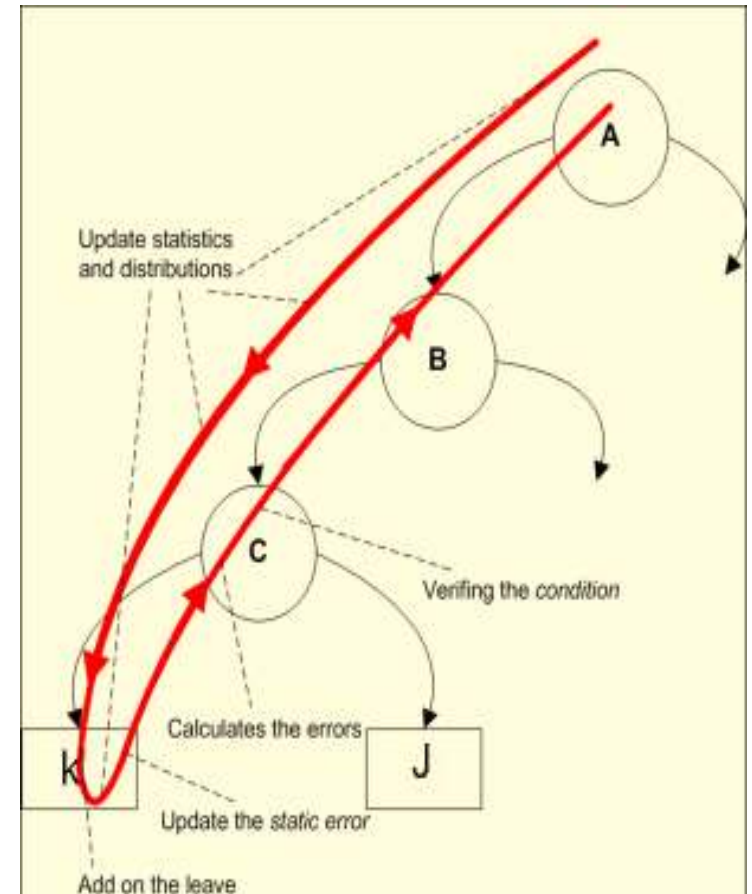


The RS Method

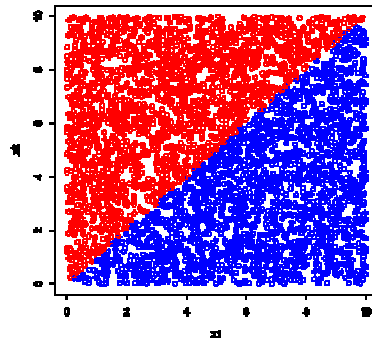
- Implemented in the VFDTc system (IDA 2006)
- For each decision node i , compute two estimates of the classification error.
 - Static error (SE_i):
 - the distribution of the error at node i ;
 - Backed up error (BUE_i):
 - the sum of the error distributions of all the descending leaves of the node i ;
 - With these two distributions:
 - we can detect the concept change,
 - by verifying the condition $SE_i \leq BUE_i$

RS Method

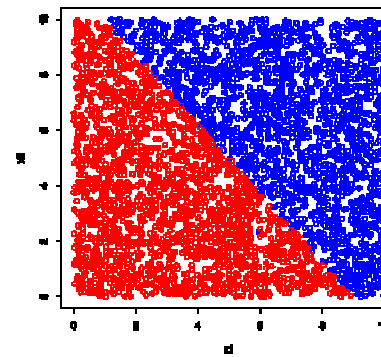
- Each new example traverses the tree from the root to a leaf
 - At the leaf: update the value of SE_i
 - It makes the opposite path, and update the values of SE_i and BUE_i for each internal node,
 - Verify the regularization condition $SE_i \leq BUE_i$:
 - If $SE_i \leq BUE_i$ then the node i is pruned to a new leaf.



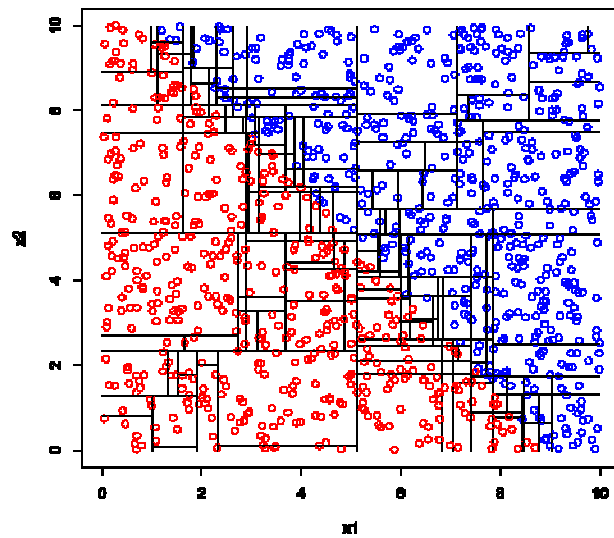
Concept 1



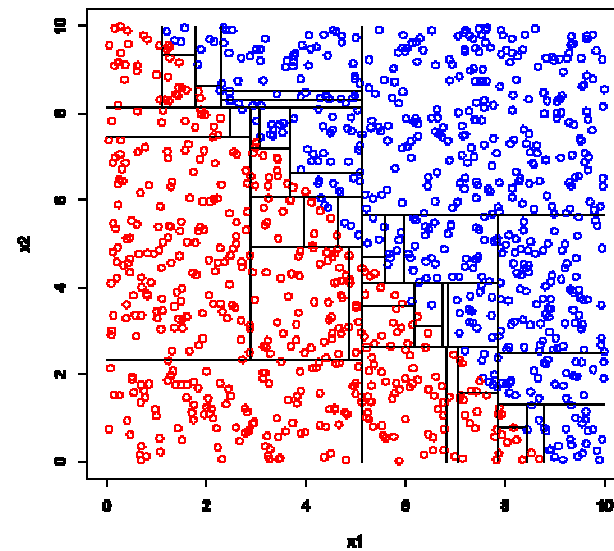
Concept 2



Without RS



With RS



ADAPTATION METHODS

Adaptation Methods

- The Adaptation model characterizes the changes in the decision model do adapt to the most recent examples.
 - Blind Methods:
 - Methods that adapt the learner at regular intervals without considering whether changes have really occurred.
 - Informed Methods:
 - Methods that only change the decision model after a change was detected. They are used in conjunction with a detection model.

Granularity of Decision Models

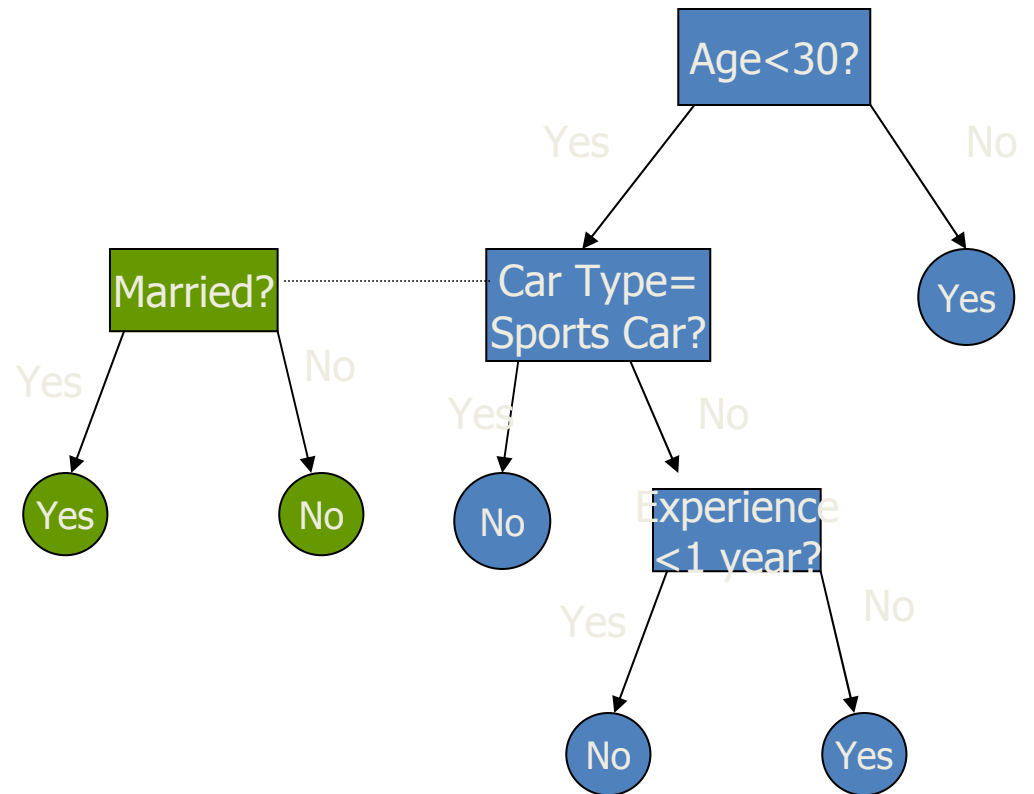
- Occurrences of drift can have impact in part of the instance space.
 - Global models: Require the reconstruction of all the decision model.
 - like naive Bayes, SVM, etc
 - Granular decision models: Require the reconstruction of parts of the decision model
 - like decision rules, decision trees

CVFDT algorithm

- *Mining time-changing data streams*, G. Hulten, L. Spencer, and P. Domingos; ACM SIGKDD 2001
- Process examples from the stream indefinitely.
 - For each example (x, y) ,
 - Pass (x, y) down to a set of leaves using HT and all alternate trees of the nodes (x, y) passes through.
 - Add (x, y) to the sliding window of examples.
- Periodically scans the internal nodes of HT.
- Start a new alternate tree when a new winning attribute is found.
 - Tighter criteria to avoid excessive alternate tree creation.
 - Limit the total number of alternate trees

Smoothly adjust to concept drift

- Alternate trees are grown the same way HT is.
- Periodically each node with non-empty alternate trees enter a testing mode.
 - M training examples to compare accuracy.
 - Prune alternate trees with non-increasing accuracy over time.
 - Replace if an alternate tree is more accurate.



Decision model management

- Model management characterizes the number of decision models needed to maintain in memory.
- The key issue here is the assumption that data generated comes from multiple distributions,
 - at least in the transition between contexts.
 - Instead of maintaining a single decision model several authors propose the use of multiple decision models.

Dynamic Weighted Majority

- A seminal work, is the system presented by Kolter and Maloof (ICDM03, ICML05).
- The Dynamic Weighted Majority algorithm (DWM) is an ensemble method for tracking concept drift.
 - Maintains an ensemble of base learners,
 - Predicts using a weighted-majority vote of these experts.
 - Dynamically creates and deletes experts in response to changes in performance.

DETECTION ALGORITHMS

Change Detection in Predictive Learning

- When there is a change in the class-distribution of the examples:
 - The actual model does not correspond any more to the actual distribution.
 - The error-rate increases
- Basic Idea:
 - Learning is a process.
 - Monitor the quality of the learning process:
 - Using Statistical Process Control techniques.
 - Monitor the evolution of the error rate.
 - Main Problems:
 - How to distinguish Change from Noise?
 - How to React to drift?

SPC-Statistical Process Control

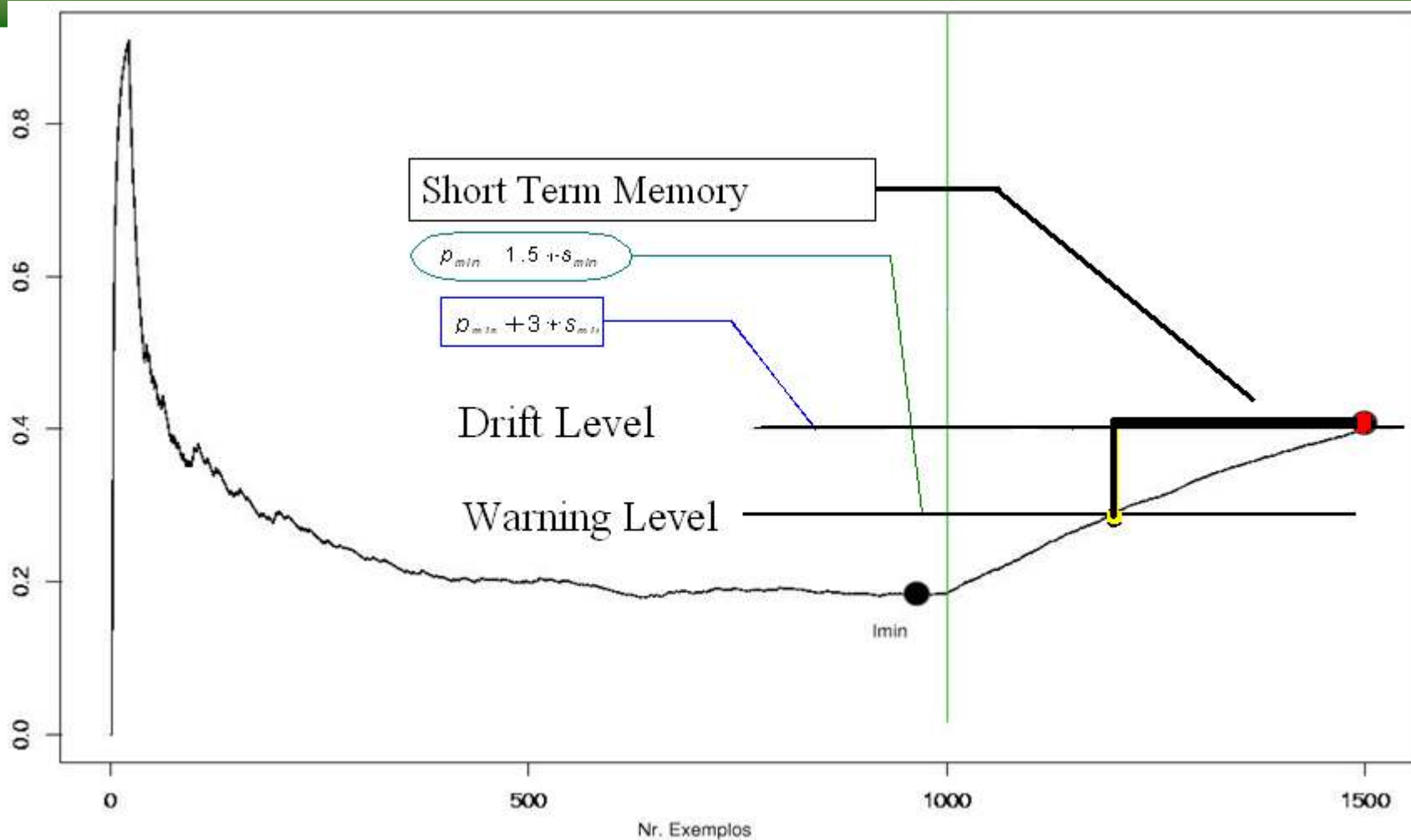
Learning with Drift Detection, Gama, Medas, Gladys, Rodrigues; SBIA-LNCS Springer, 2004.

- Suppose a sequence of examples in the form $\langle \mathbf{x}_i, y_i \rangle$
- The actual decision model classifies each example in the sequence
 - In the 0-1 loss function, predictions are either True or False
 - The predictions of the learning algorithm are sequences:
T, F, T, F, T, F, T, T, T, F, ...
 - The Error is a random variable from Bernoulli trials.
 - The Binomial distribution gives the general form of the probability of observing a F:
 - $p_i = (\#F/i)$
 - $s_i = \sqrt{p_i(1-p_i)/i}$ where i is the number of trials.

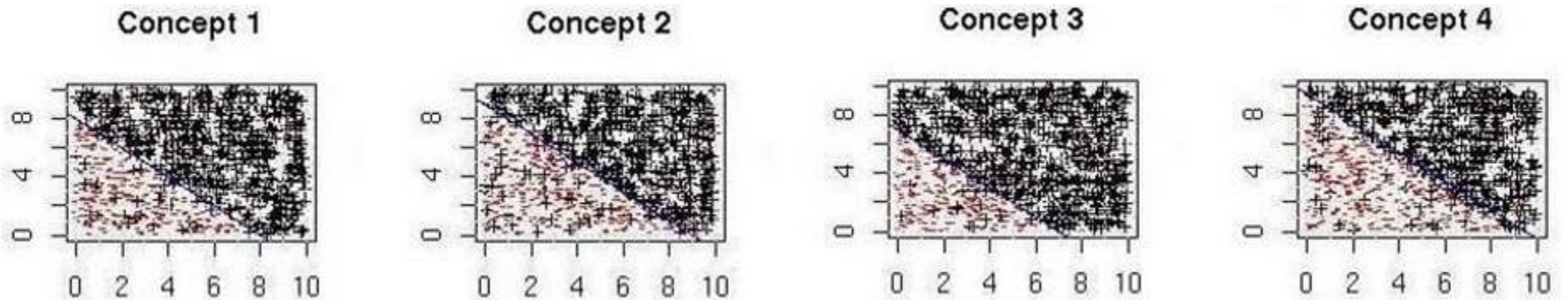
Statistical Processing Control Algorithm

- Maintains two registers:
 - P_{\min} and S_{\min} such that $P_{\min} + S_{\min} = \min(p_i + s_i)$
 - Minimum of the error rate taking into account the variance of the estimator.
- At example j : the error of the learning algorithm will be
 - **Out-control** if $p_j + s_j > p_{\min} + \alpha S_{\min}$
 - **In-control** if $p_j + s_j < p_{\min} + \beta S_{\min}$
 - **Warning Level**: if $p_{\min} + \alpha S_{\min} > p_j + s_j > p_{\min} + S_{\min}$
- The constants α and β depend on the desired confidence level.
 - Admissible values are $\alpha = 2$ and $\beta = 3$.

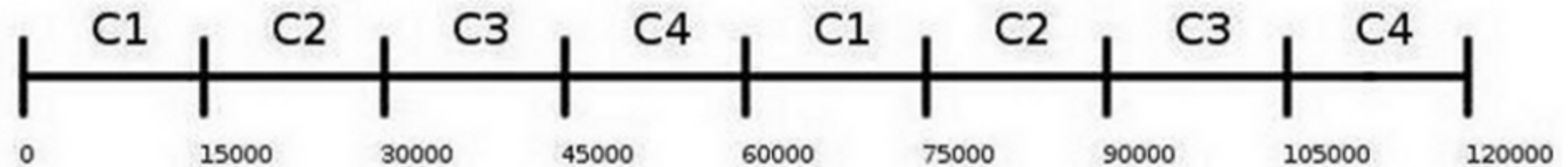
Statistical Processing Control



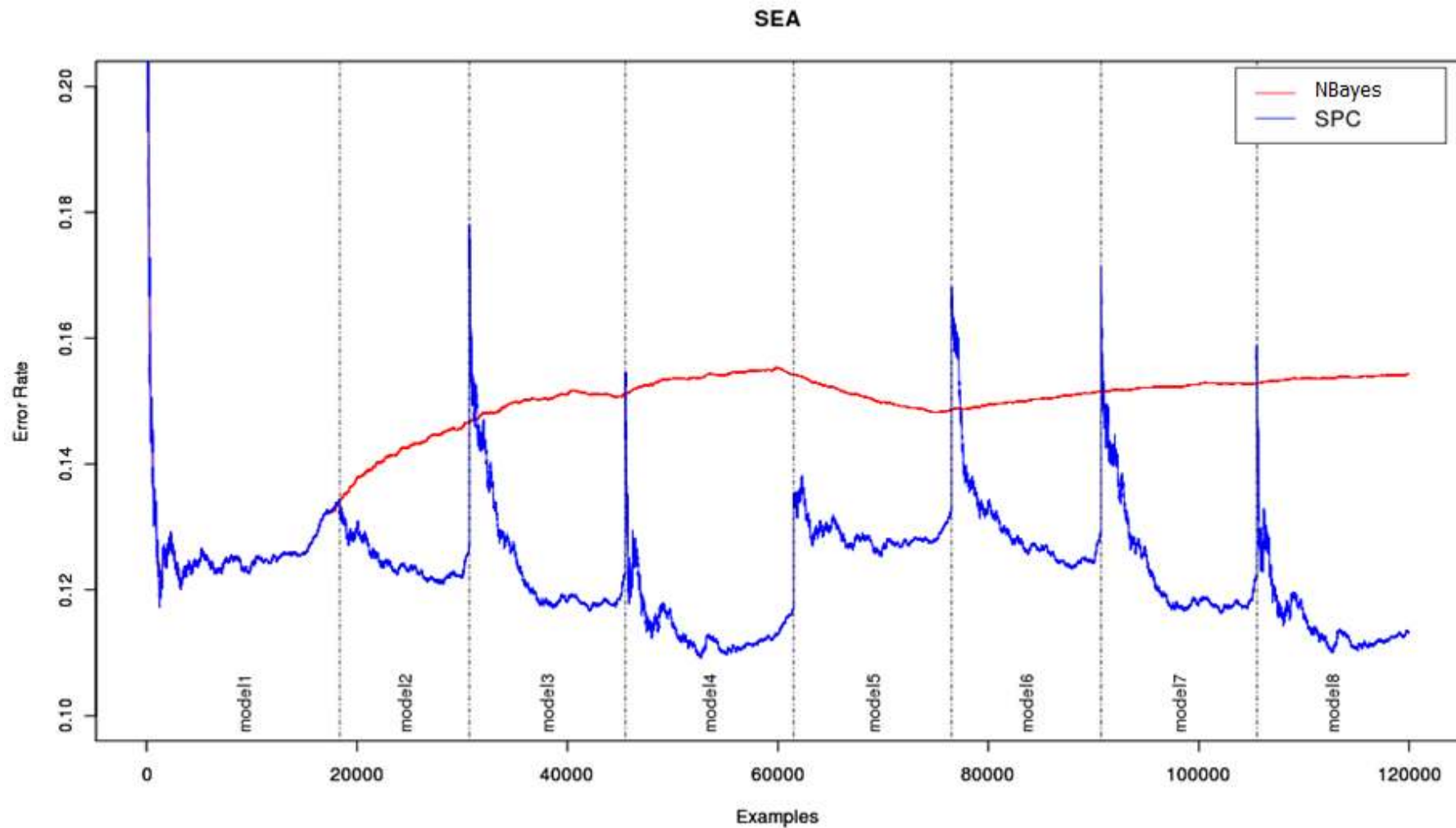
Illustrative Example: SEA Concepts



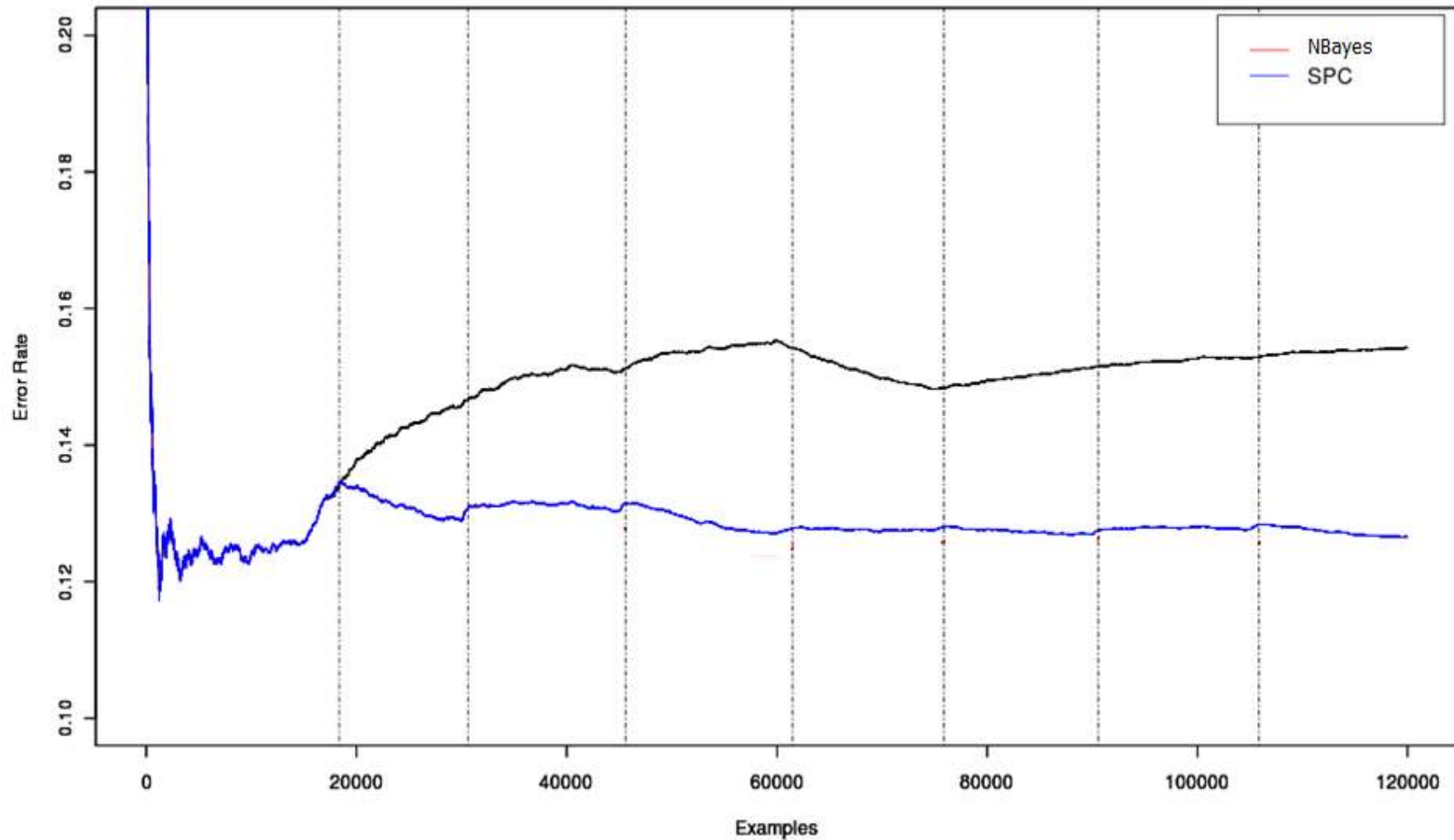
- The stream:



The Individual Error of the Models



System Error



Sequential Analysis

- Sequential analysis developed a set of methods for monitoring change detection:
 - Sequential Probability Ratio Test - SPRT algorithm, D.Wald, *Sequential Tests of Statistical Hypotheses*. Annals of Mathematical Statistics 16 (2): 117–186; 1945
 - Cumulative Sums - CUSUM: E. Page, *Continuous Inspection Scheme*. Biometrika 41,1954

CUSUM Algorithms

The CUSUM test is used to detect significant increases (or decreases) in the successive observations of a random variable \mathbf{x}

- Detecting significant increases:
 - $\mathbf{g}_0 = 0$
 - $\mathbf{g}_t = \max(0; \mathbf{g}_{t-1} + (\mathbf{x}_t - \alpha))$
- The decision rule is:
 - if $\mathbf{g}_t > \lambda$ then alarm and $\mathbf{g}_t = 0$.
- Detecting decreases:
 - $\mathbf{g}_0 = 0$
 - $\mathbf{g}_t = \min(0; \mathbf{g}_{t-1} + (\mathbf{x}_t - \alpha))$
- The decision rule is:
 - if $\mathbf{g}_t < -\lambda$ then alarm and $\mathbf{g}_t = 0$.

Page-Hinckley Test

- The PH test is a sequential adaptation of the detection of an abrupt change in the average of a Gaussian signal.
 - It considers a cumulative variable m_T , defined as the cumulated difference between the observed values and their mean till the current moment:

$$m_{t+1} = \sum_1^t (x_t - \bar{x}_t + \alpha)$$

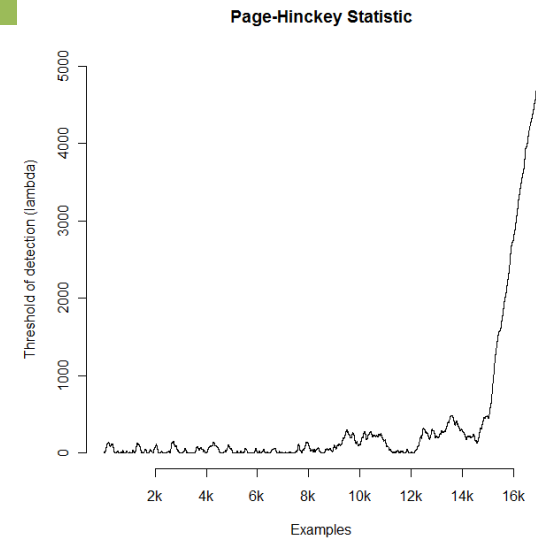
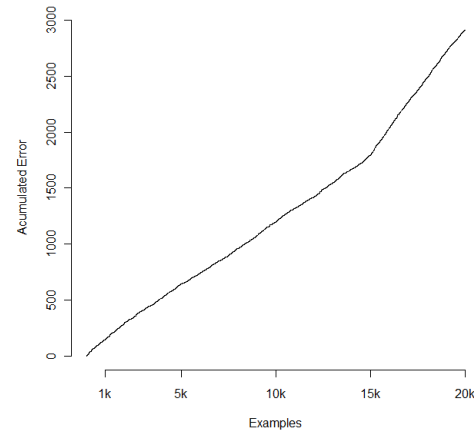
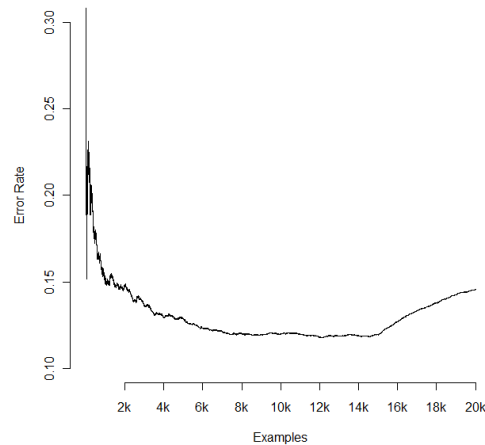
- where $\bar{x} = 1/t \sum_{l=1}^t x_l$ and
- α corresponds to the magnitude of changes that are allowed.

The Page-Hinckley Test

$$m_{t+1} = \sum_1^t (x_t - \bar{x}_t + \alpha)$$

- The minimum value of m_t is also computed:
 - $M_T = \min(m_t ; t = 1, \dots, T)$.
- The test monitors the difference between M_T and m_T :
 - $PH_T = m_T - M_T$.
- When this difference is greater than a given threshold (λ) we alarm a change in the distribution.

Illustrative Example



- The left figure plots the on-line error rate of a learning algorithm.
- The center plot is the accumulated on-line error. The slope increases after the occurrence of a change.
- The right plot presents the evolution of the PH statistic.

Page-Hinckley Test

- The PH_t test is memoryless, and its accuracy depends on the choice of parameters α and λ .
 - Both parameters are relevant to control the trade-off between earlier detecting true alarms by allowing some false alarms.
- The threshold λ depends on the admissible false alarm rate.
 - Increasing λ will entail fewer false alarms, but might miss some changes.

Algorithm Adaptive Window-ADWIN

- *Learning from Time-Changing Data with Adaptive Windowing*, A.Bifet, R.Gavalda (SDM'07)
- Whenever two “large enough” subwindows of W exhibit “distinct enough” averages,
 - we can conclude that the corresponding expected values are different, and
 - the older portion of the window is dropped

ADWIN0: ADAPTIVE WINDOWING ALGORITHM

```
1 Initialize Window  $W$ 
2 for each  $t > 0$ 
3   do  $W \leftarrow W \cup \{x_t\}$  (i.e., add  $x_t$  to the head of  $W$ )
4   repeat Drop elements from the tail of  $W$ 
5     until  $|\hat{\mu}_{W_0} - \hat{\mu}_{W_1}| < \epsilon_{cut}$  holds
6     for every split of  $W$  into  $W = W_0 \cdot W_1$ 
7   output  $\hat{\mu}_W$ 
```

Algorithm ADWIN

- The value of ϵ_{cut} for a partition $W_0 \cdot W_1$ of W is computed as follows:
 - Let n_0 and n_1 be the lengths of W_0 and W_1 and
 - Let n be the length of W , so $n = n_0 + n_1$.
 - Let η_{W_0} and η_{W_1} be the averages of the values in W_0 and W_1 , and W_0 and W_1 their expected values.
 - To obtain totally rigorous performance guarantees we define:

$$m = \frac{1}{1/n_0 + 1/n_1} \text{ (harmonic mean of } n_0 \text{ and } n_1\text{),}$$

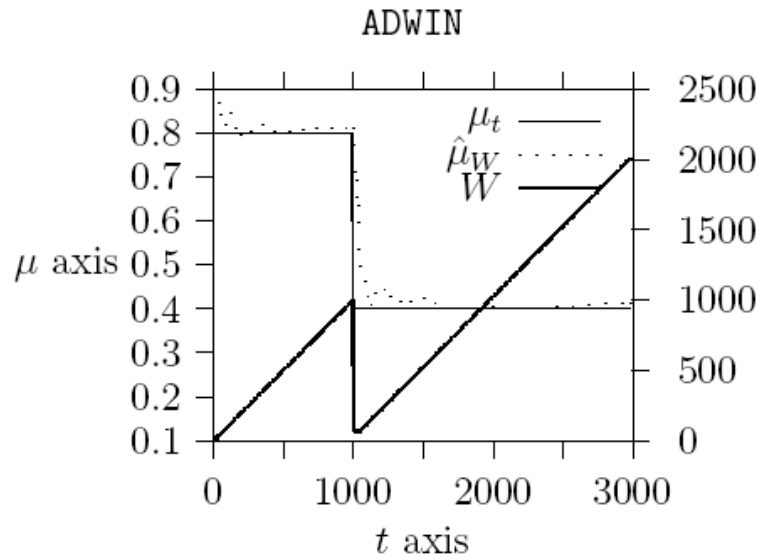
$$\delta' = \frac{\delta}{n}, \text{ and } \epsilon_{\text{cut}} = \sqrt{\frac{1}{2m} \cdot \ln \frac{4}{\delta'}}.$$

Characteristics of Algorithm ADWIN

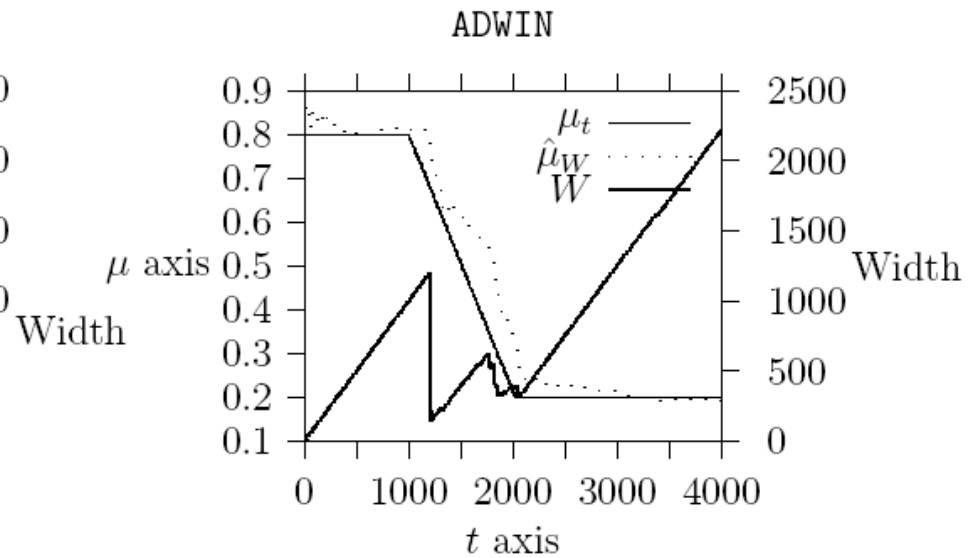
- False positive rate bound:
 - If η_t remains constant within W , the probability that ADWIN shrinks the window at this step is at most δ .
- False negative rate bound.
 - Suppose that for some partition of W in two parts W_0W_1 (where W_1 contains the most recent items) we have:
 $|\eta_{W_0} - \eta_{W_1}| > 2^* \epsilon_{\text{cut}}$.
Then with probability $1 - \delta$ ADWIN shrinks W to W_1 , or shorter.

Illustrative Examples

Abrupt Change



Gradual Change



ADWIN – Compressing Space

- Using exponential histograms
- Requires $O(\log W)$ space and time
- It can provide exact counts for all the subwindows

Example: Counting 1's in a bit stream

A new 1 arrives:

1010101	101	11	1	1	1
Content: 4	2	2	1	1	1
Capacity: 7	3	2	1	1	1

Compress:

There are 3 buckets of size 1:

1010101	101	11	11	1
Content: 4	2	2	2	1
Capacity: 7	3	2	2	1

Compress again:

There are 3 buckets of size 2:

1010101	10111	11	1
Content: 4	4	2	1
Capacity: 7	5	2	1

Remove buckets when a drift is detect:

10111	11	1
Content: 4	2	1
Capacity: 5	2	1

Summary

Dimensions of Analysis:

- **Data management**
 - Characterize the information stored in memory to maintain a decision model consistent with the actual state of the nature
- **Detection Methods**
 - Characterizes the techniques and mechanisms for explicit drift detection
- **Adaptation methods**
 - Characterizes the changes in the decision model do adapt to the most recent examples.
- **Decision model management**
 - Characterize the number of decision models needed to maintain in memory.

Block 3: evaluation and MOA demo

The goals:

- to present the main training and evaluation principles of adaptive learning approaches
- to present a tool that is designed for building and evaluating adaptive learners:

– an open-source software

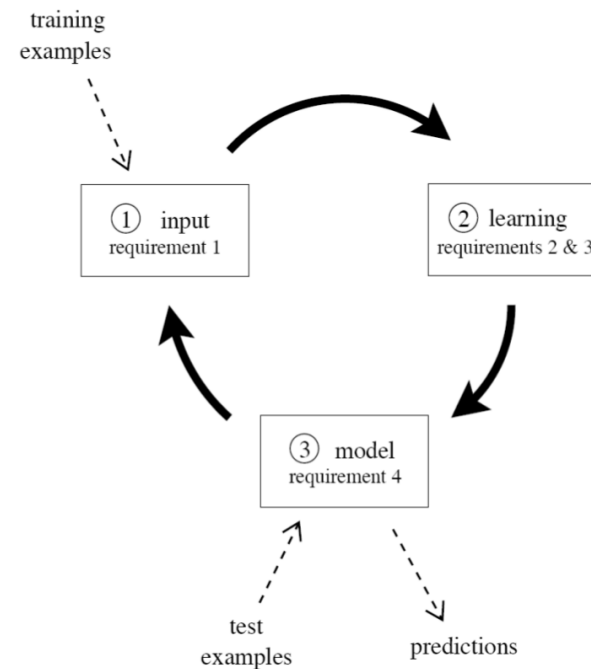


Adaptive Learning Method Evaluation

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

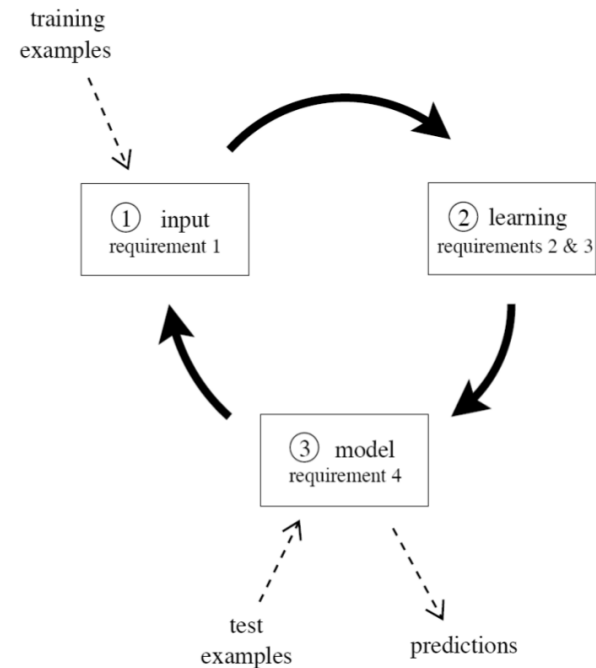
Streaming Setting

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



Streaming Evaluation

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



Holdout Evaluation

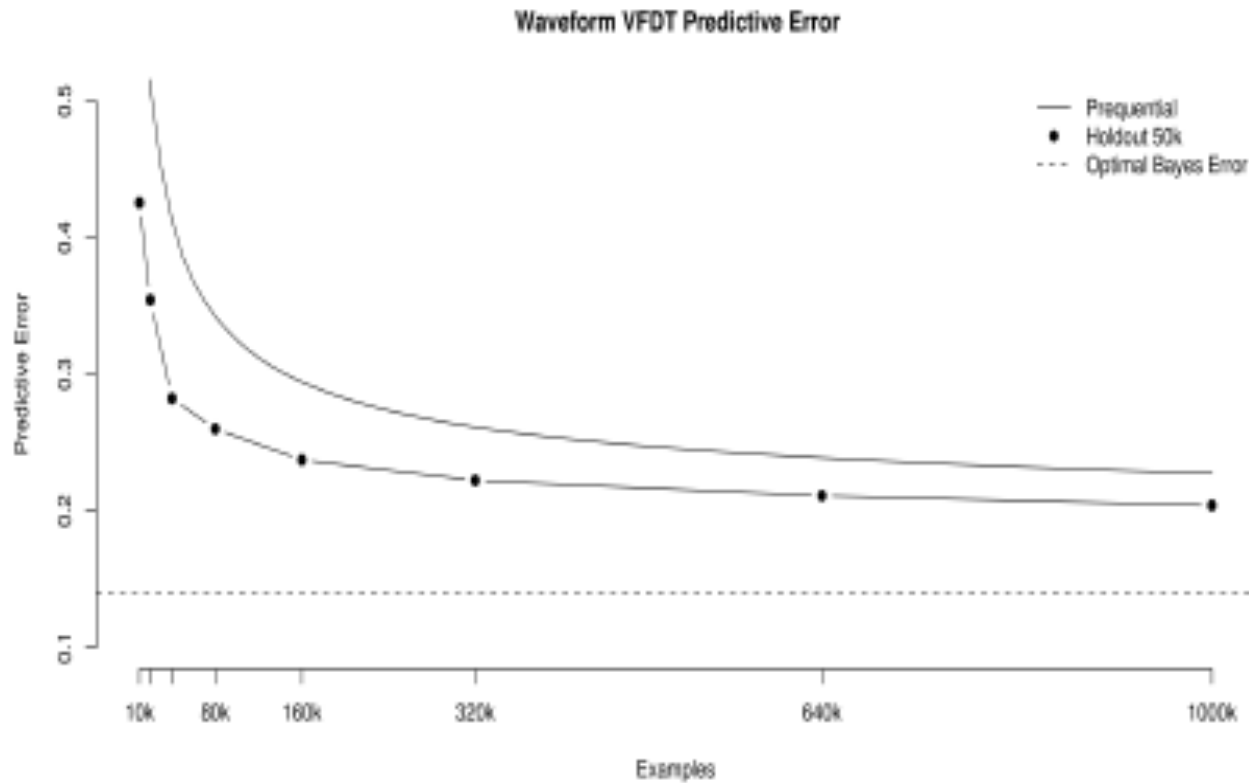
- Holdout an independent test set
- Apply the current decision model to the test set, at regular time intervals
- The loss estimated in the holdout is an unbiased estimator

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.

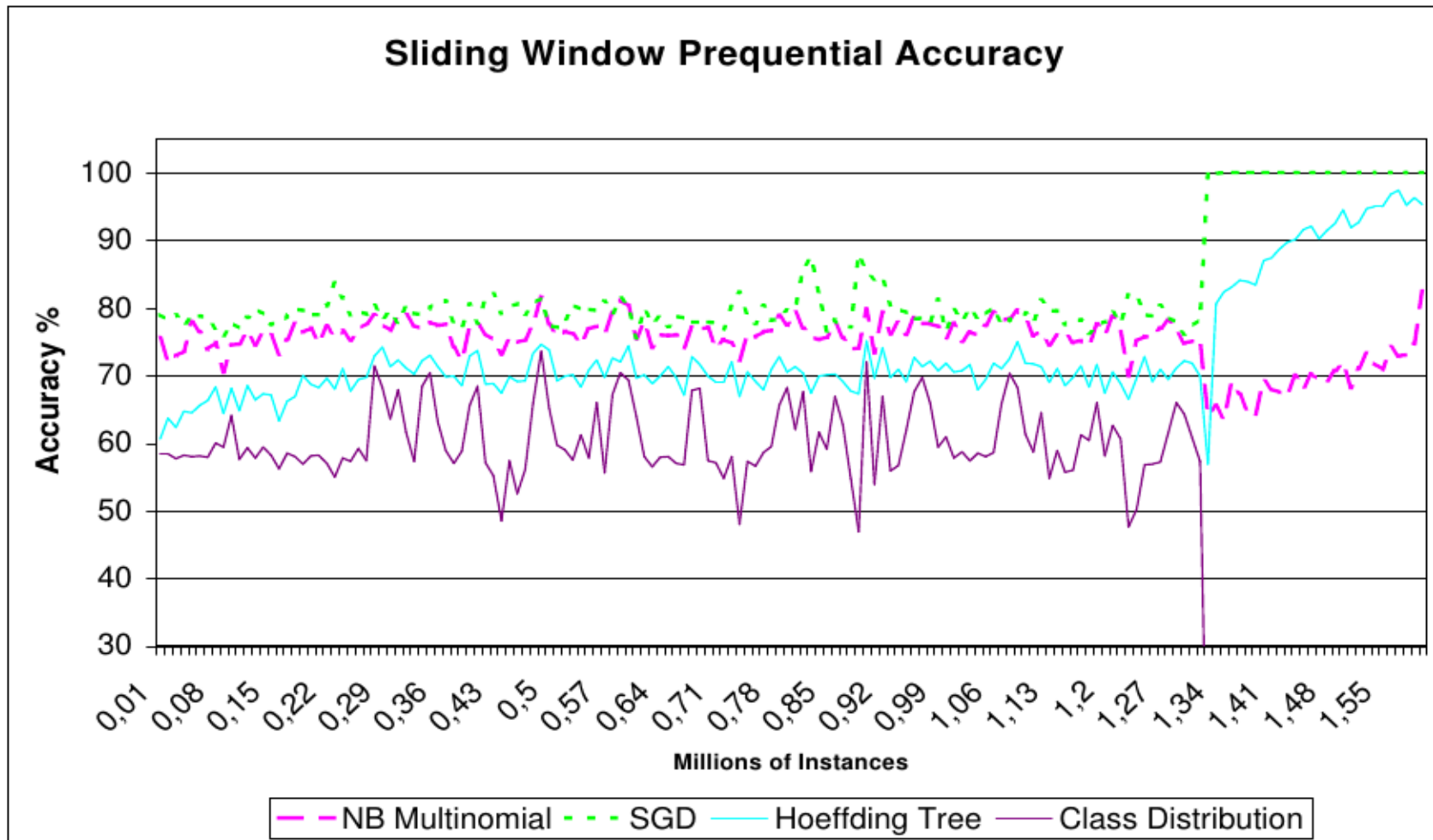
$$S = \sum_{i=1}^n L(y_i, \hat{y}_i).$$

Prequential Evaluation

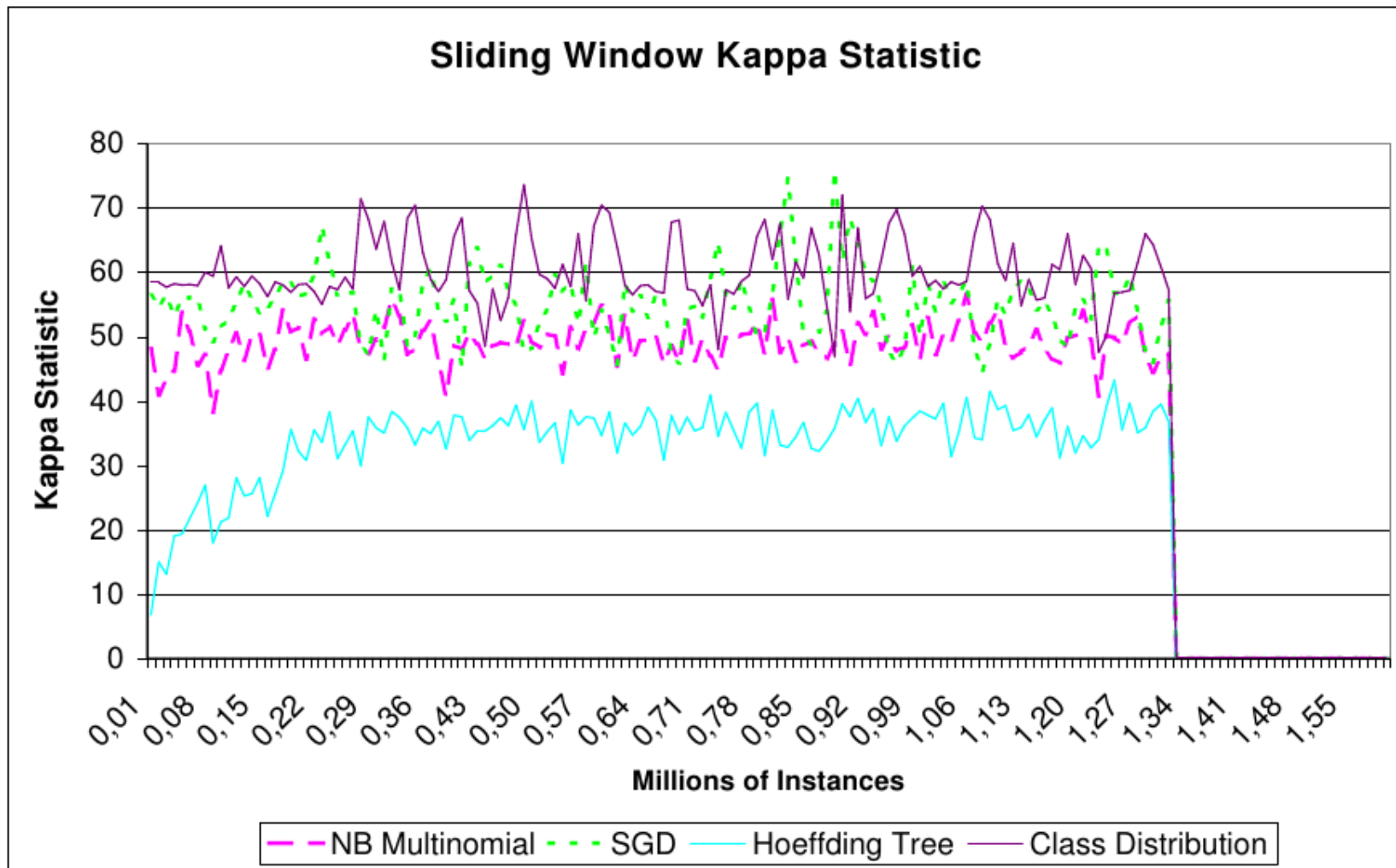


Issues in Evaluation of Stream Learning Algorithms, Joao Gama, Raquel Sebastiao, Pedro Pereira Rodrigues, KDD 2009

Twitter Example



Twitter Example



Kappa Statistic

	Predicted Class +	Predicted Class -	Total
Correct Class +	75	8	83
Correct Class -	7	10	17
Total	82	18	100

Simple confusion matrix example

	Predicted Class +	Predicted Class -	Total
Correct Class +	68.06	14.94	83
Correct Class -	13.94	3.06	17
Total	82	18	100

Confusion matrix for chance predictor

Kappa Statistic

- p_0 : classifier's prequential accuracy
- p_c : probability that a chance classifier makes a correct prediction.

$$\kappa \text{ statistic} = (p_0 - p_c) / (1 - p_c)$$

- $\kappa = 1$ if the classifier is always correct
- $\kappa = 0$ if the predictions coincide with the correct ones as often as those of the chance classifier

RAM Hours

	Accuracy	Time	Memory	
Classifier A	70%	100	20	
Classifier B	80%	20	40	

RAM Hours

- Time and Memory in one measure
- Every GB of RAM deployed for 1 hour
- Inspired by Cloud Computing Rental Cost Options

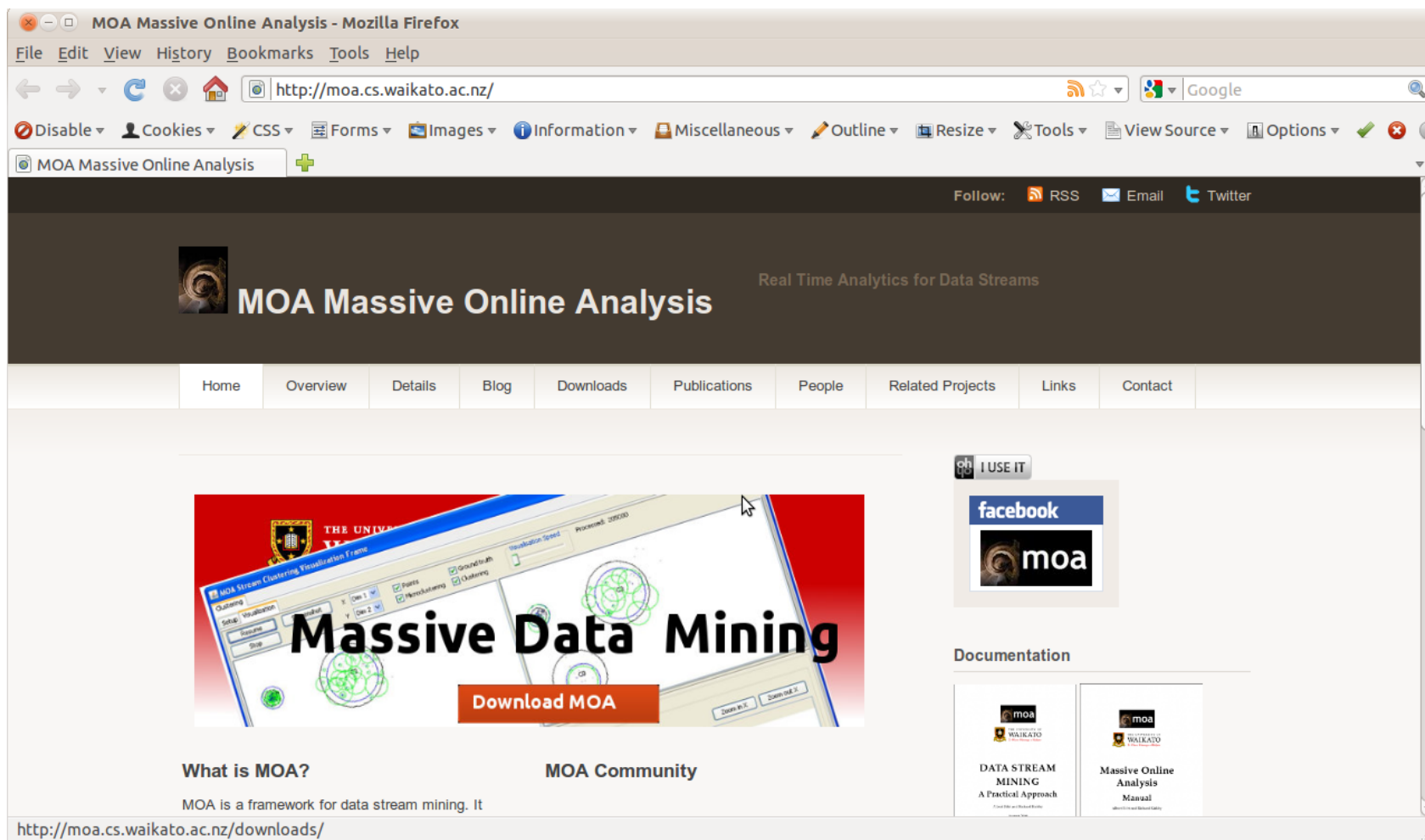
RAM Hours

	Accuracy	Time	Memory	RAM-Hours
Classifier A	70%	100	20	2000
Classifier B	80%	20	40	800

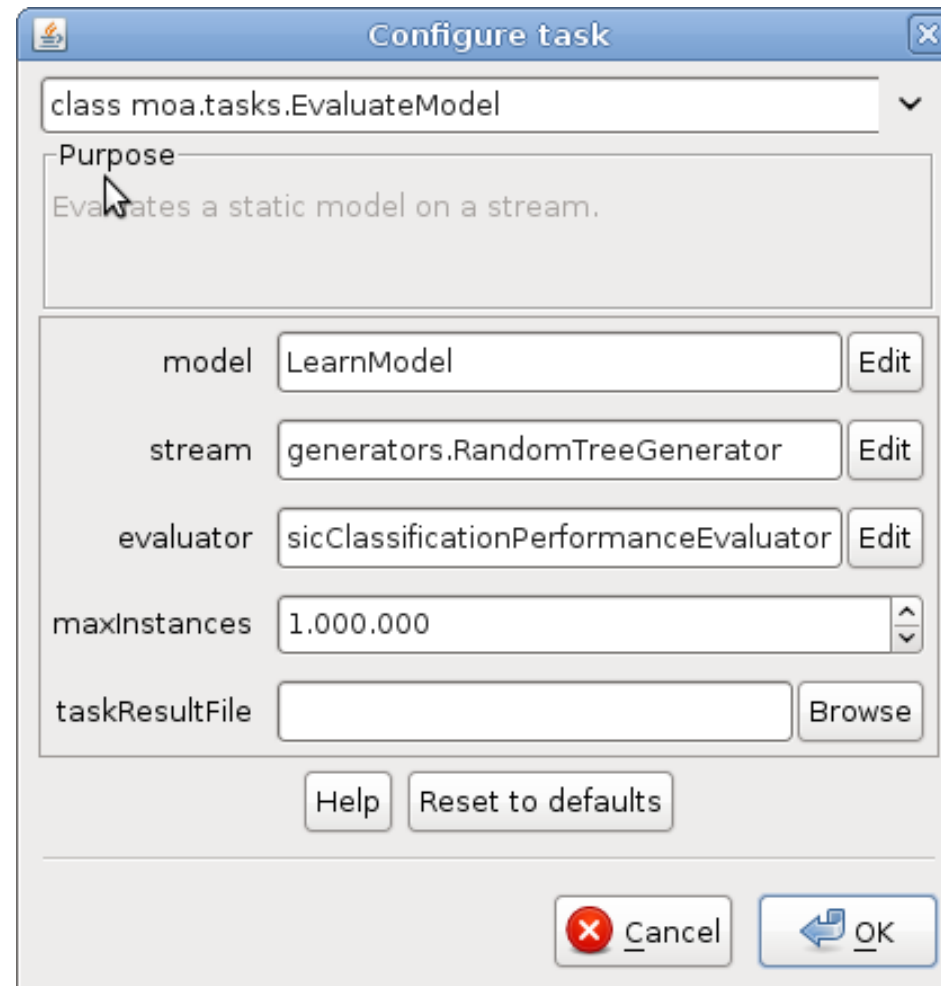


MOA Software

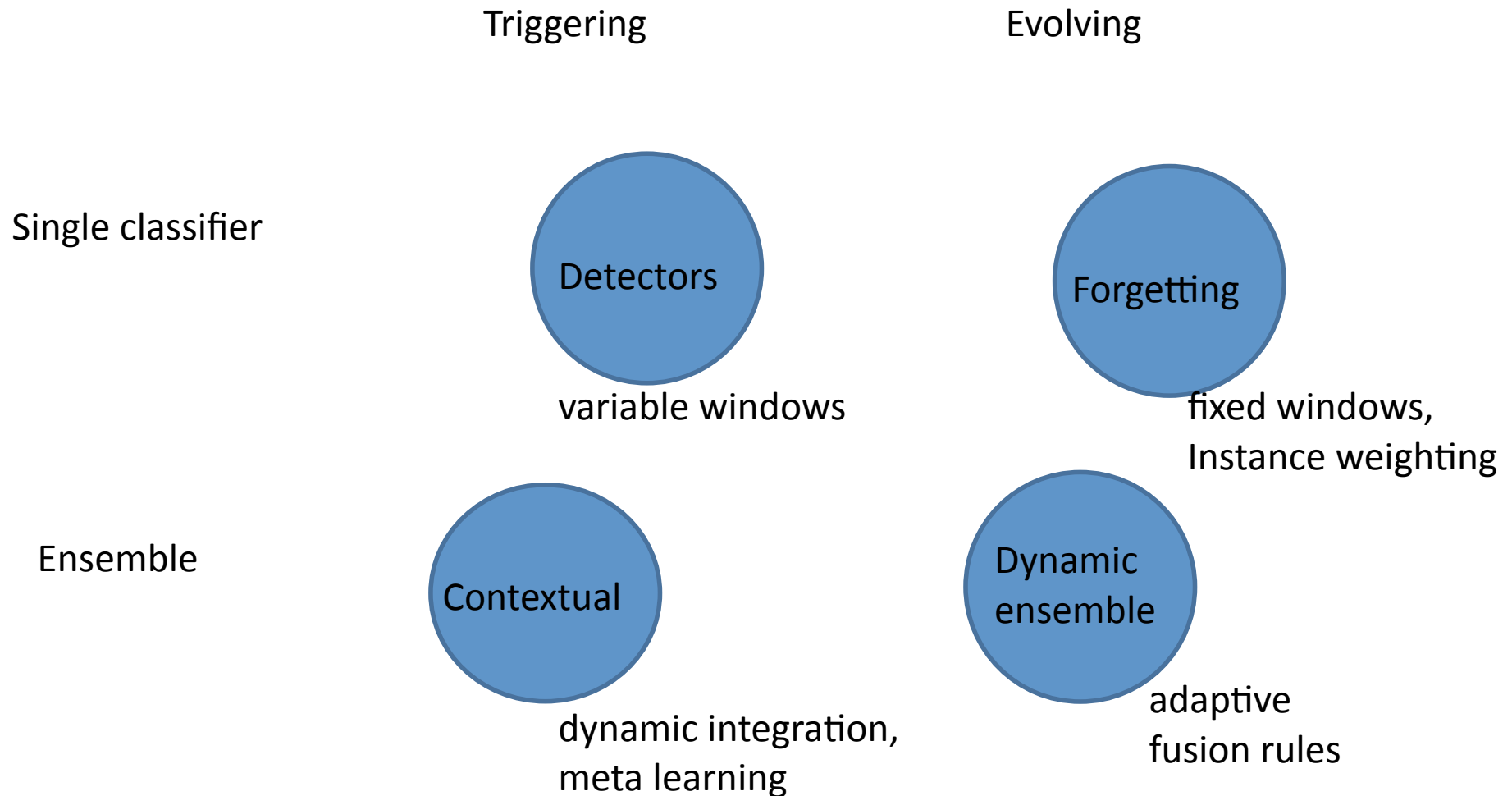
Introduction



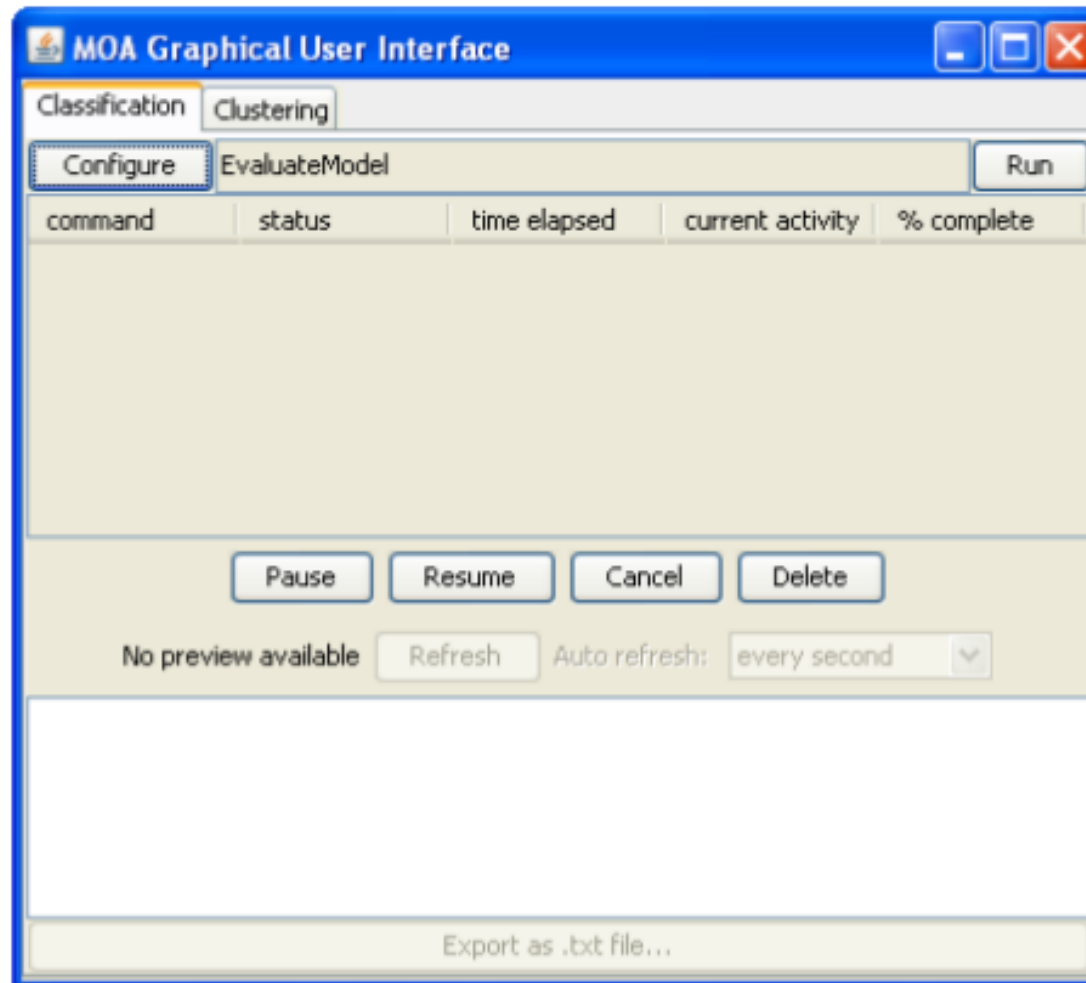
Evaluation



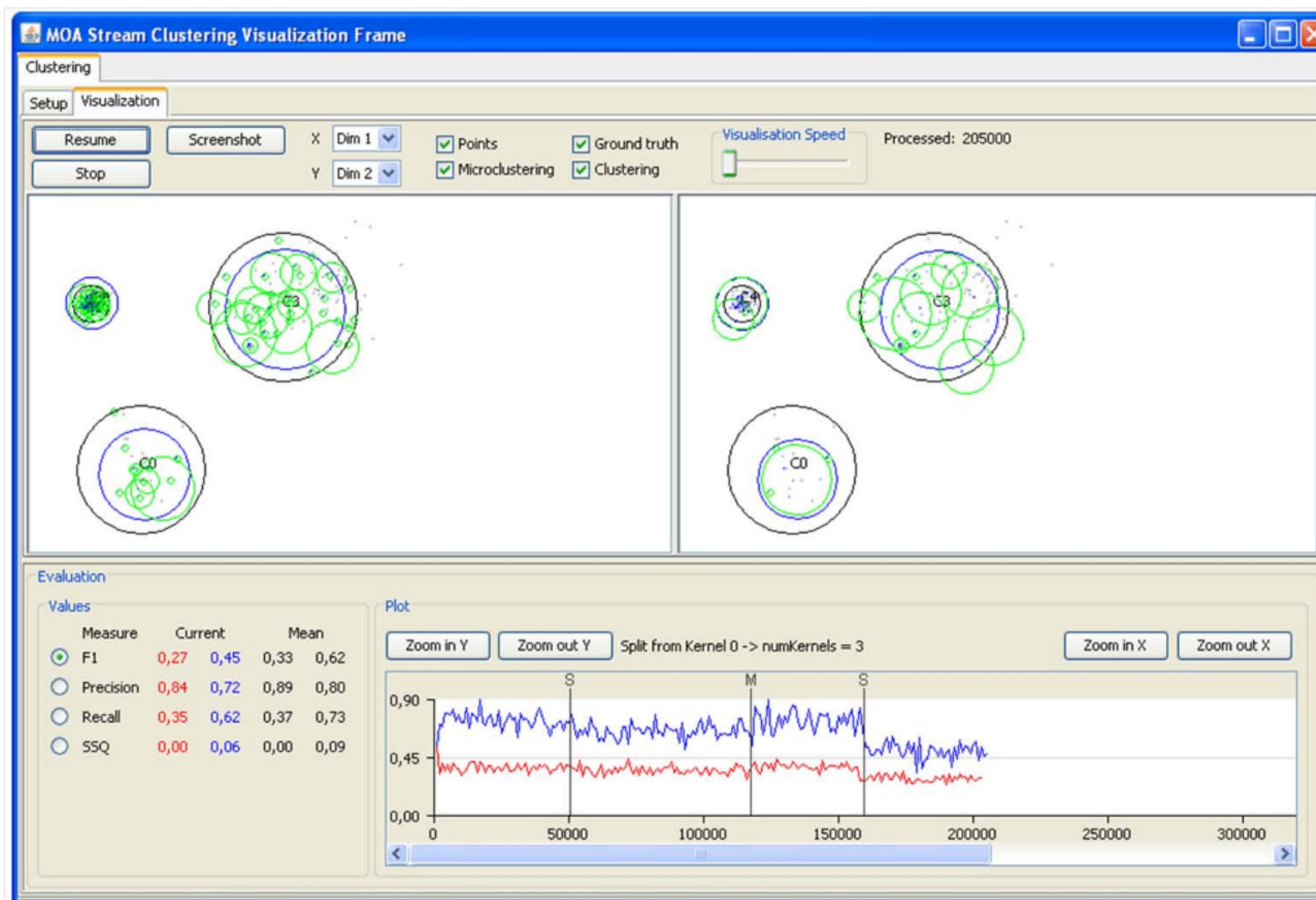
Techniques to Handle Concept Drift



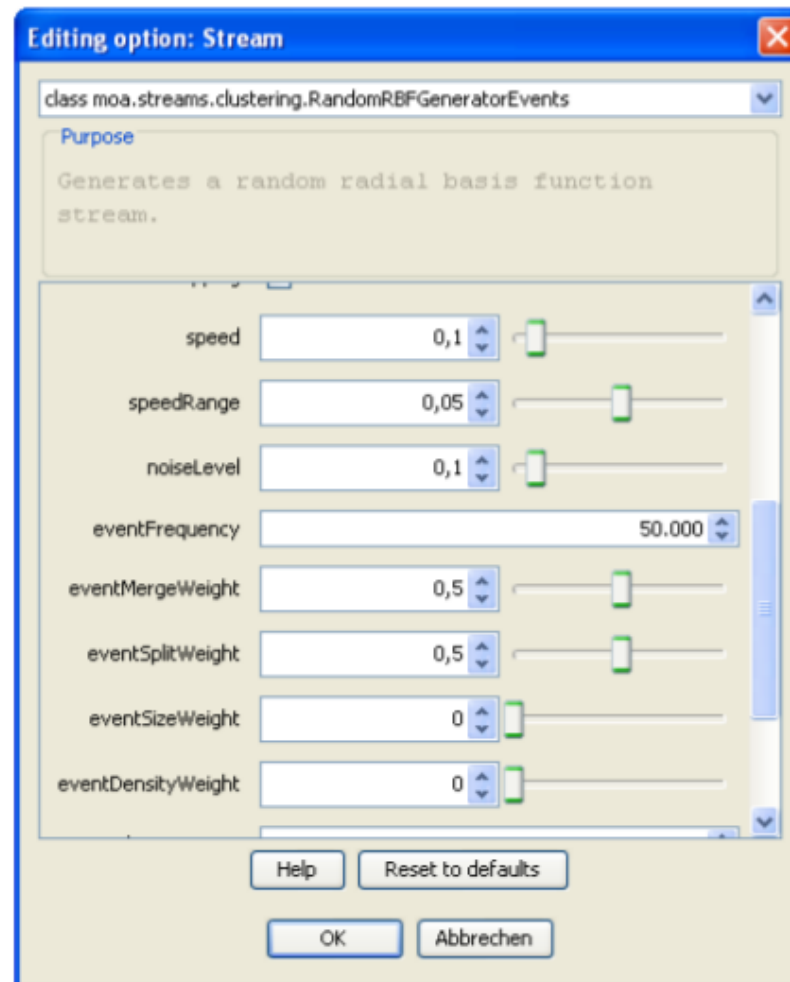
Classification



Clustering



Graphical Interface



Command Line

EvaluatePeriodicHeldOutTest

- ```
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

# How to build a learner

- `void resetLearningImpl ()`
- `void trainOnInstanceImpl  
(Instance inst)`
- `double[] getVotesForInstance  
(Instance i)`



# Summary Block 3

- Streaming Evaluation
- Accuracy using Prequential Evaluation
- Kappa Statistic and RAM-Hours
- MOA: open source software for data streams
  - easy to use and extend

# Block 4: An application perspective

# Block 4: Outline

Goal I: provide researchers dealing with CD related problems a **reference framework** for:

- positioning their work with respect to the applicability of the techniques in practice,
- helping young researchers to avoid “green aliens” problem in motivating their work,
- better see new research directions driven by *challenges* and *opportunities* of certain types of **real applications** where CD matters

Goal II: present highlights of a few case studies

# Classical situation in early DM

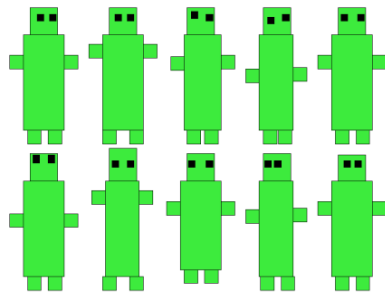
- Problem is believed to be important, but
  - real data is hard to get due to privacy or proprietary (but also laziness) reasons
  - or the solution is too “generic” such that it is not applicable in practice
  - or the solution addresses a tiny problem in the overall automated decision making process
- At extreme
  - There is no *real* application in mind and it is substituted by an *invented* real application (“green aliens”)

# As a consequence

- A lot of research is done on small/toy UCI datasets and simulated/artificial data
- Exaggerated in concept drift area
  - Moving hyperplane and SEA concepts (artificially generated) benchmarks and UCI datasets “adjusted” for the needs of researchers (artificially introduced drift) have been used in many publications on CD

# “Green Aliens” as an extreme case:

- Inventing non-existing applications and challenges
- See the parody paper by Richard D. Arami
  - “Novel Efficient Automated Robust Eye Detection System for Green Aliens”

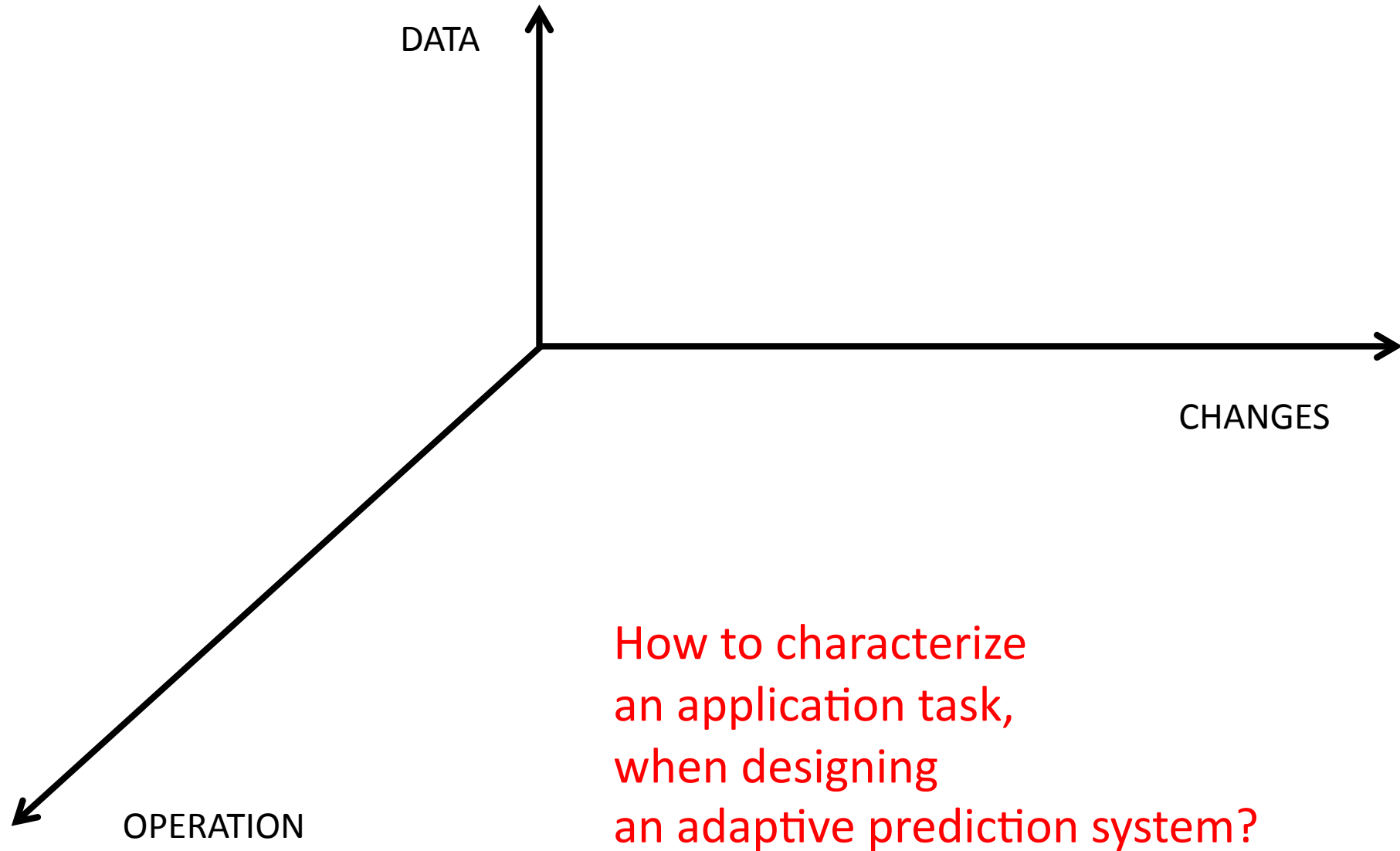


[https://sites.google.com/site/richardaramid/green\\_alien.pdf](https://sites.google.com/site/richardaramid/green_alien.pdf)

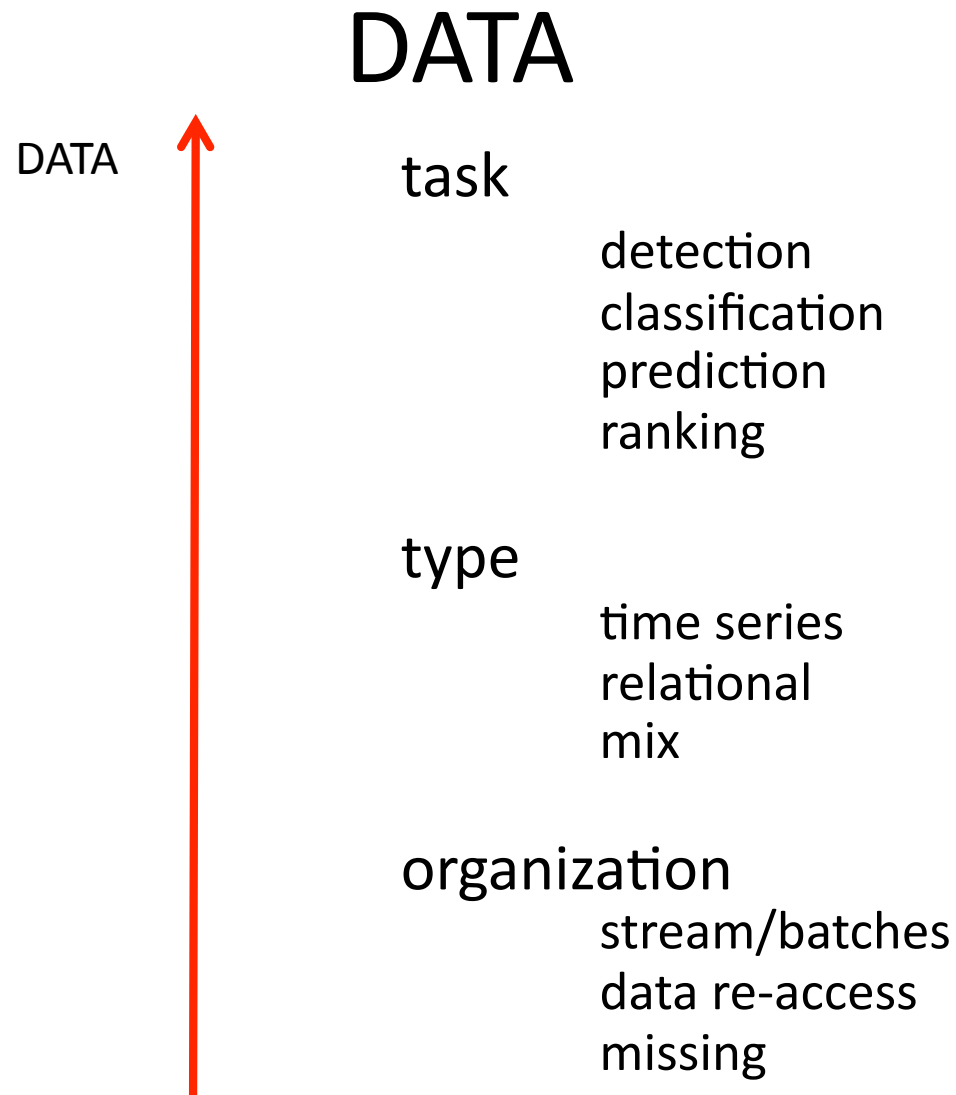
# Block 4: Outline

- Reference framework for the categorization of applications in which CD matters
  - Type of learning/mining task, available data and labels, and expectations on type/nature of CD
- Case studies
  - Peculiarities for different kind or applications
  - What challenges and opportunities they bring

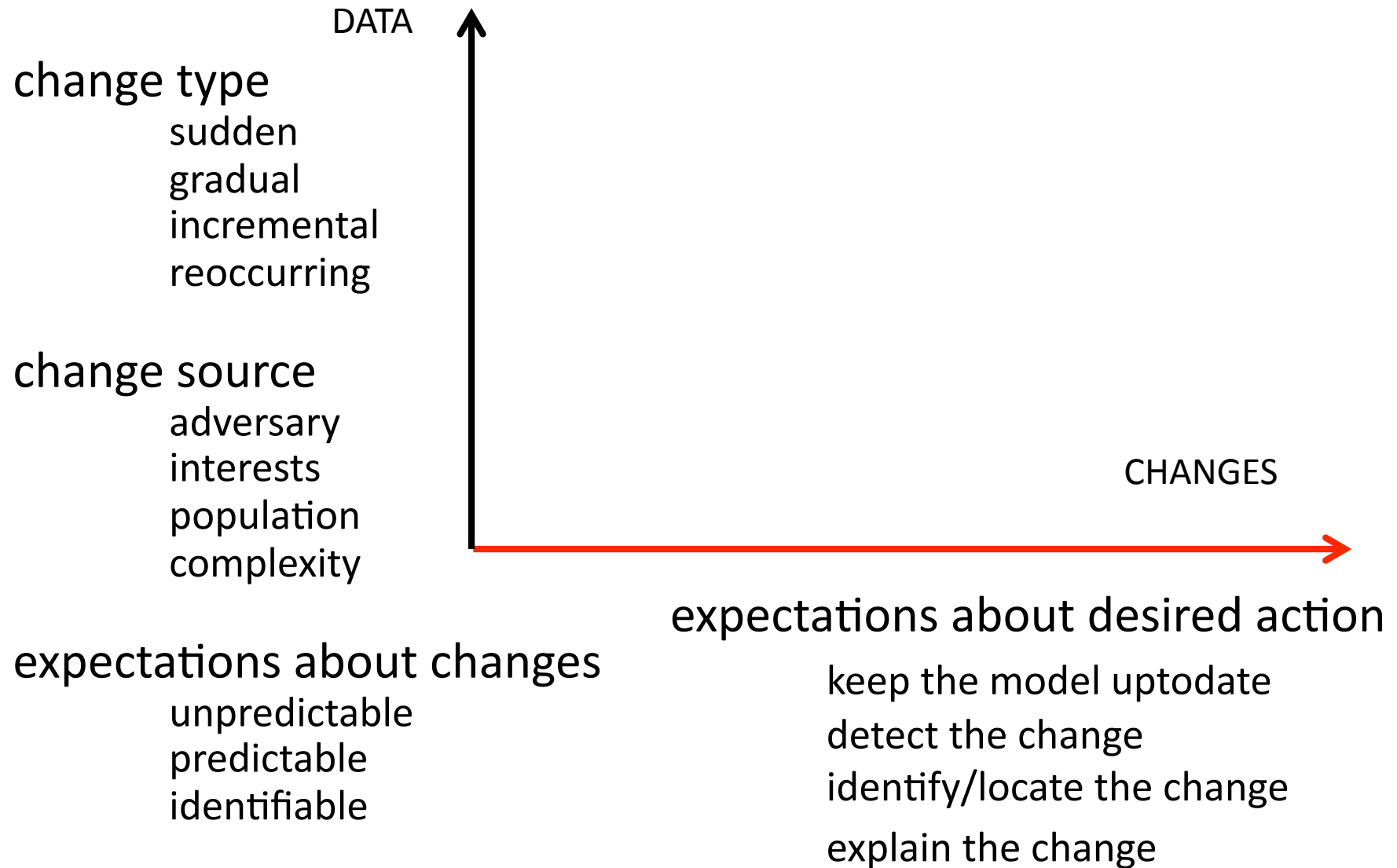
# Application tasks

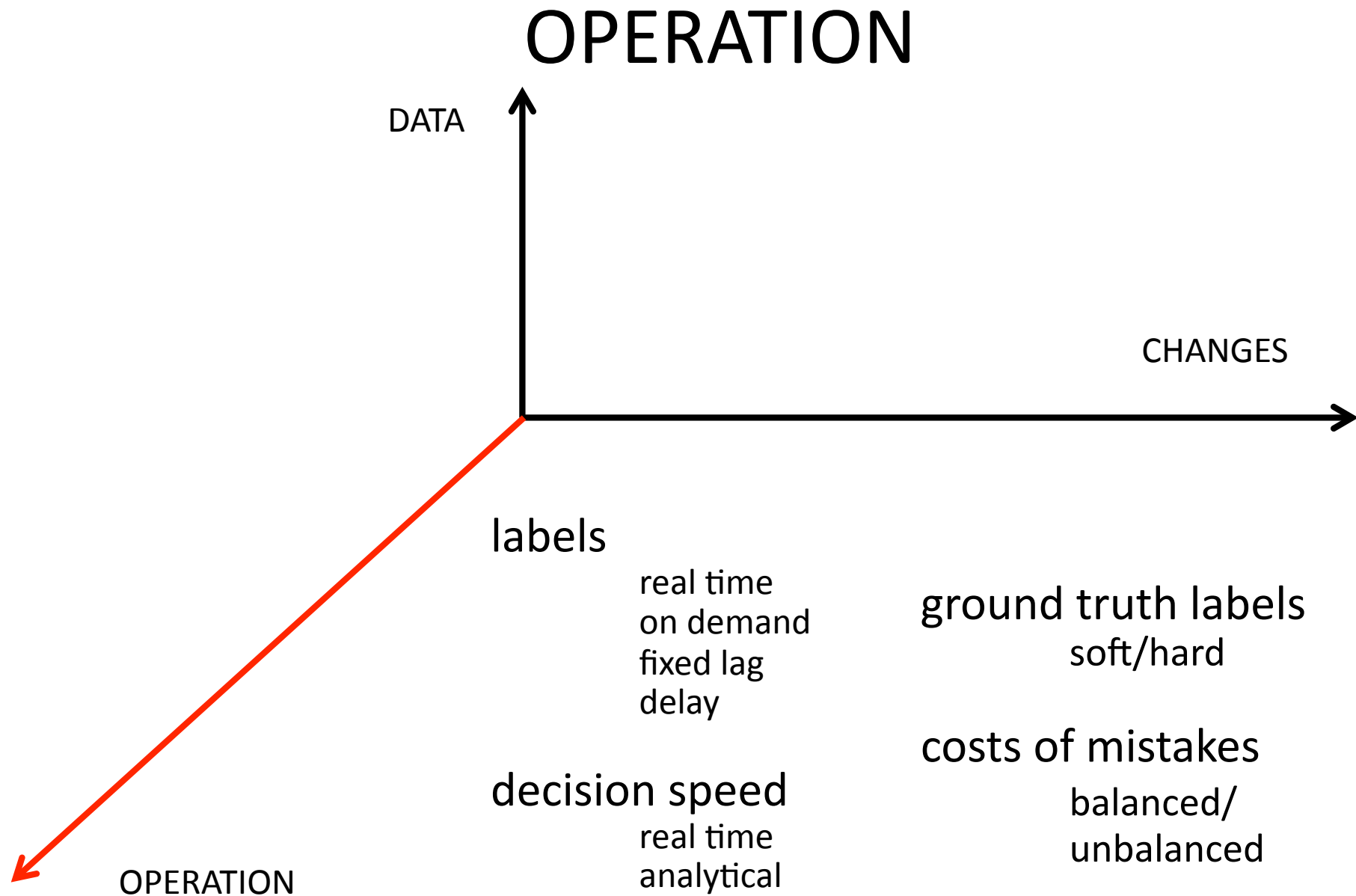


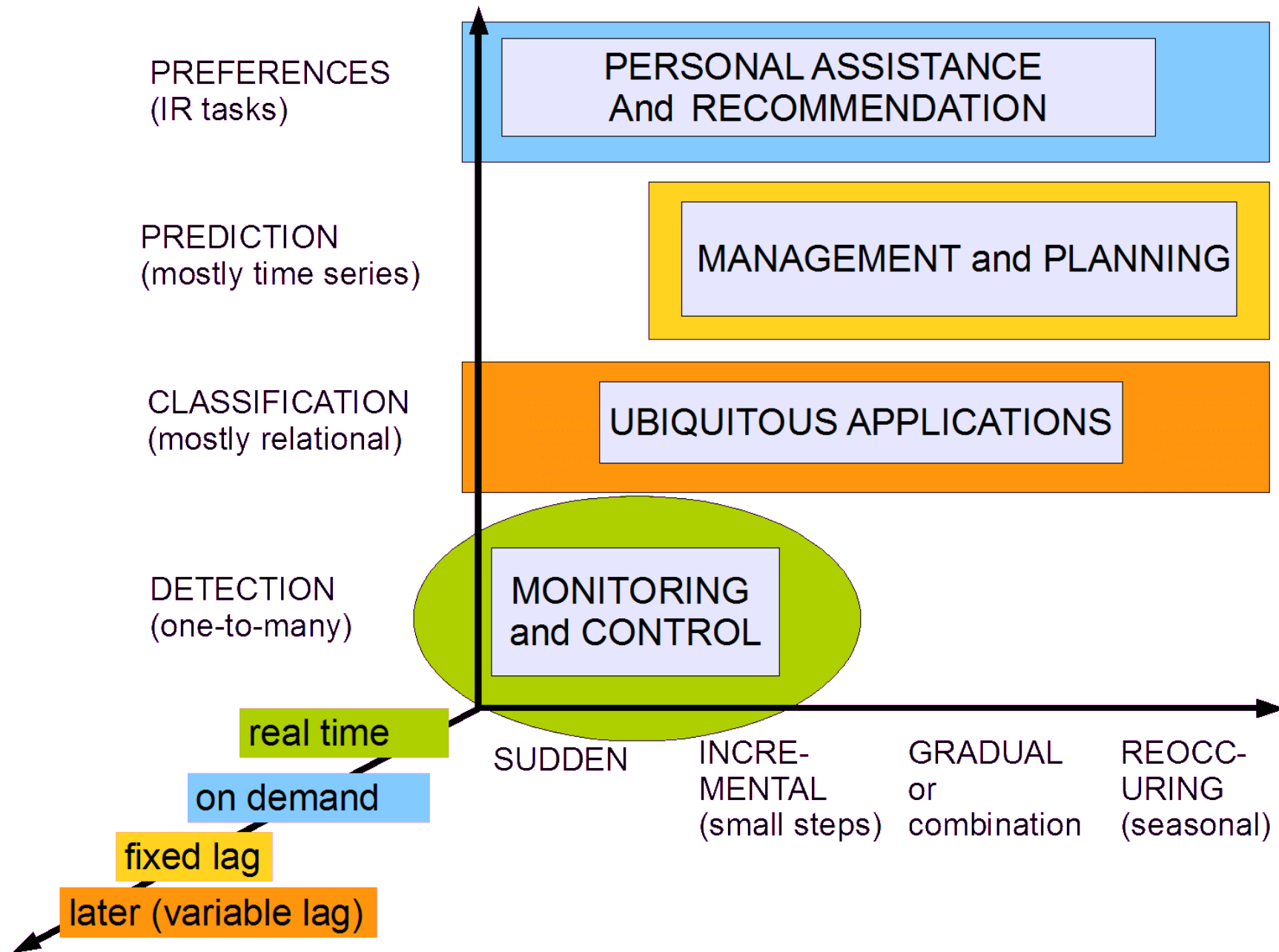




# CHANGE

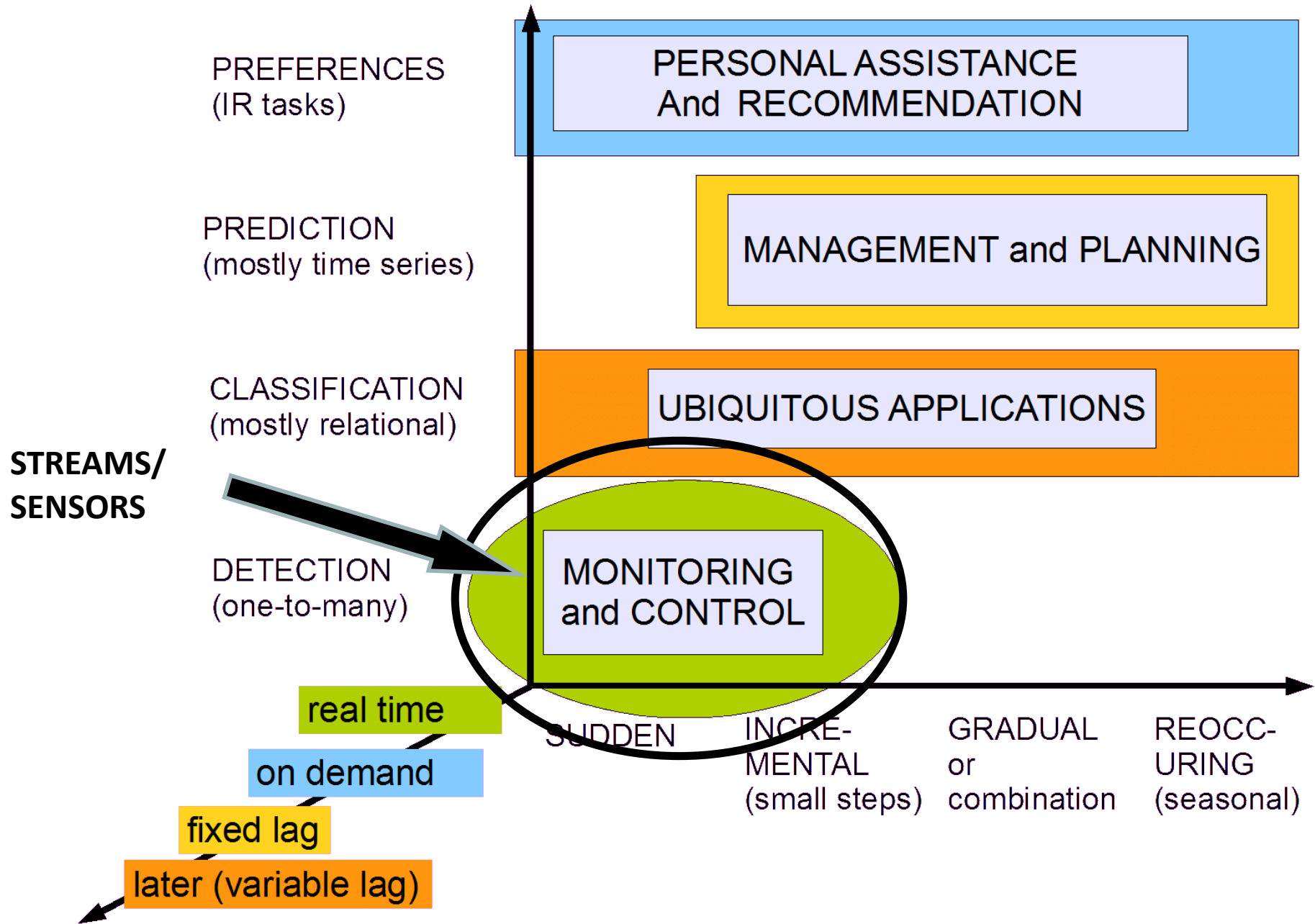




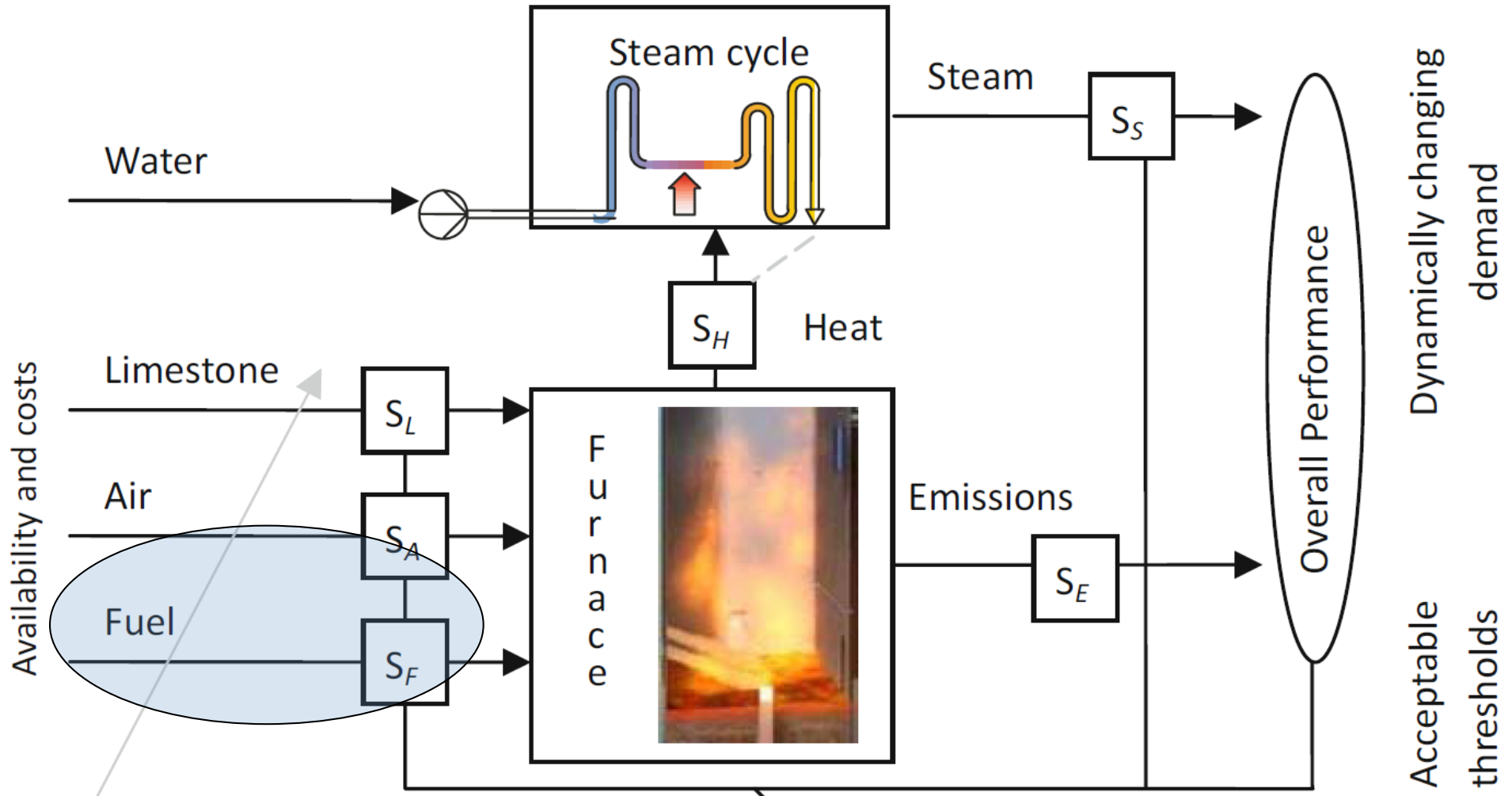


# Landscape of applications

| <i>Types of apps</i><br><i>Industries</i>                                                                               | Monitoring/<br>control                                                              | Personal assistance/<br>personalization                                                                     | Management<br>and planning                                                  | Ubiquitous<br>applications                                    |
|-------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------|---------------------------------------------------------------|
| Security, Police                                                                                                        | Fraud detection,<br>insider trading<br>detection,<br>adversary actions<br>detection | -----                                                                                                       | Crime volume<br>prediction                                                  | Authentica-<br>tion,<br>Intrusion<br>detection                |
| Finance, Banking,<br>Telecom, Credit<br>Scoring, Insurance,<br>Direct Marketing,<br>Retail, Advertising, e-<br>Commerce | Monitoring &<br>management of<br>customer<br>segments,<br>bankruptcy<br>prediction  | Product or service<br>recommendation,<br>including<br>complimentary                                         | Demand<br>prediction,<br>response rate<br>prediction,<br>budget<br>planning | Location<br>based<br>services,<br>related ads,<br>mobile apps |
| Education (higher,<br>professional, children,<br>e-Learning)<br>Entertainment, Media                                    | Gaming the<br>system,<br>Drop out<br>prediction                                     | Music, VOD, movie,<br>learning object<br>recommendation,<br>adaptive news<br>access, personalized<br>search | Player-<br>centered game<br>design,<br>learner-<br>centered<br>education    | Virtual reality,<br>simulations                               |

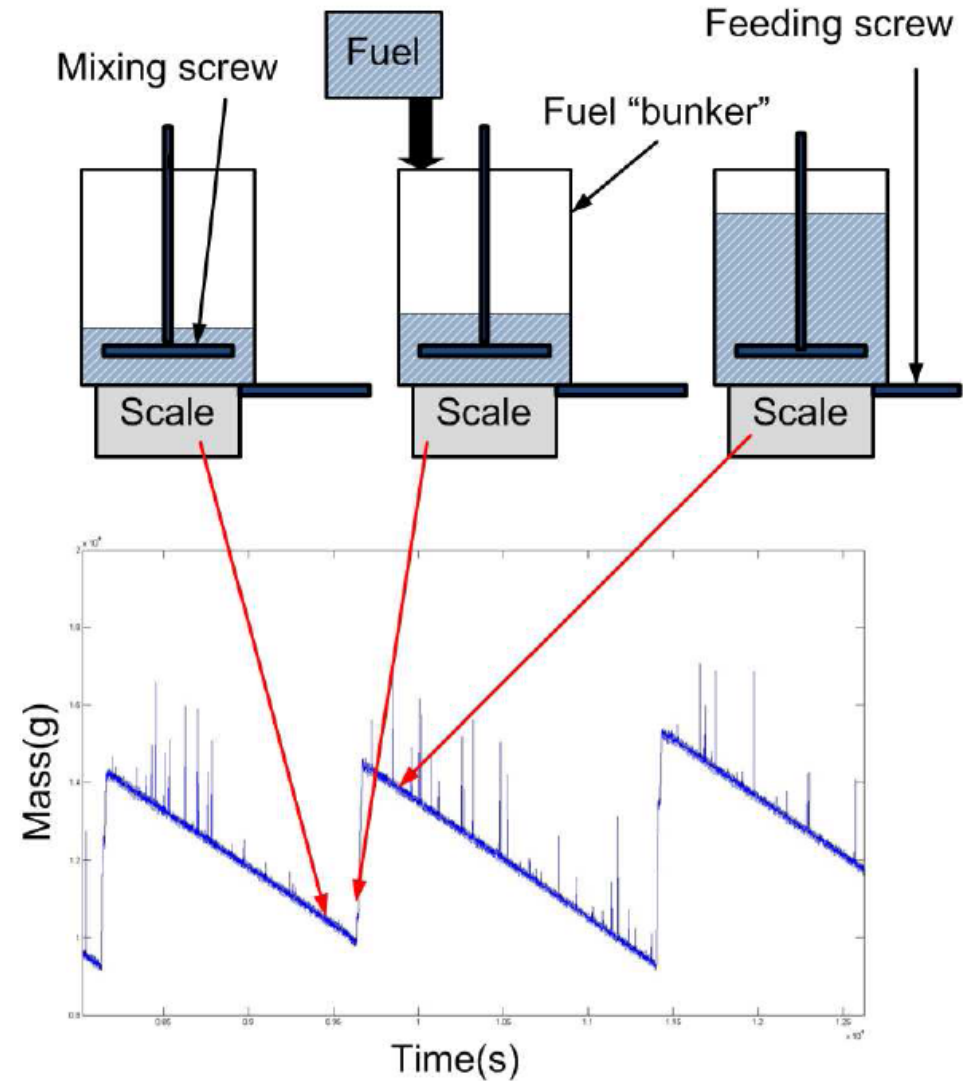


# CFB Boiler Optimization



# Online Mass Flow Prediction

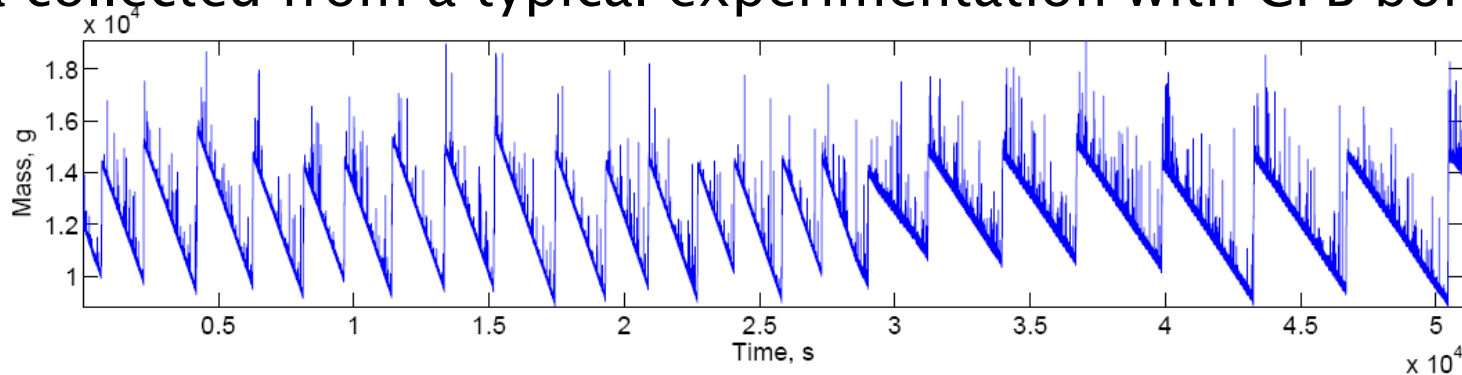
- fluctuation in the signal of the scales  $\Rightarrow$  no reliable online data can be obtained from the mass flow of fuel to the boiler
- predict the *actual* mass flow based on the sensor measurements
- two phases: feed and burn





# Online Mass Flow Prediction

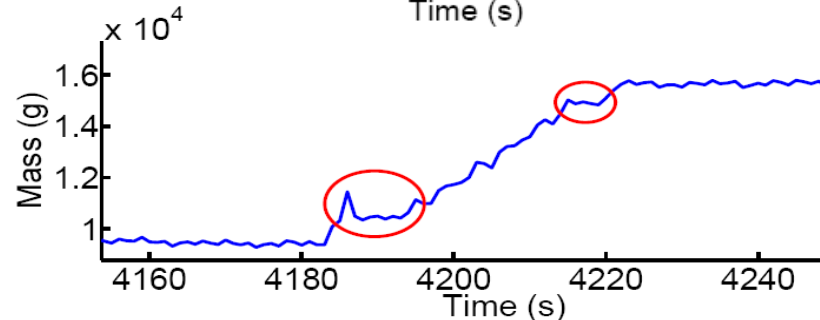
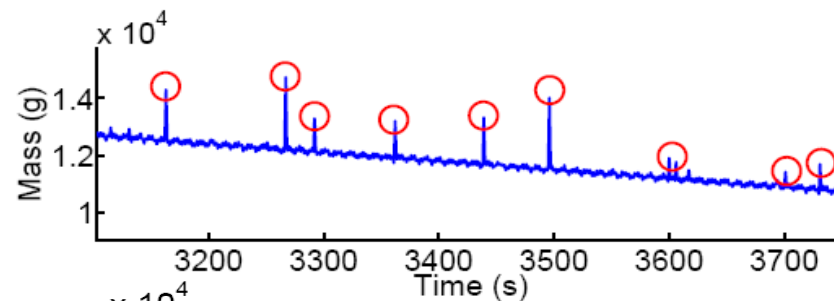
data collected from a typical experimentation with CFB boiler




asymmetric nature  
of the outliers

short consumption  
periods within  
feeding stages

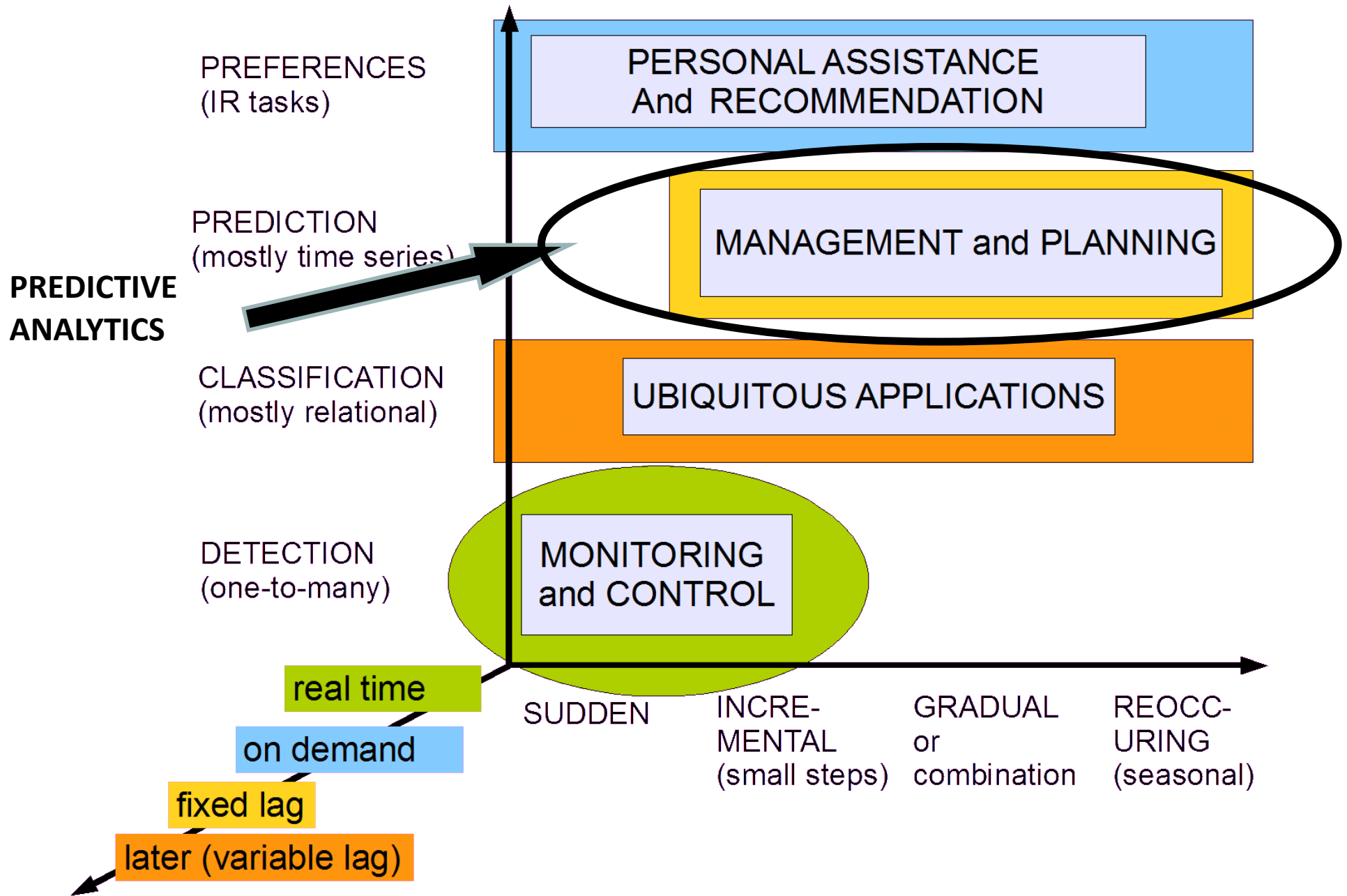
Pechenizkiy et al. 2009



# Solution – use domain knowledge

- Adaptive learning approach
    - Detect a change and cut the training window
  - Challenges
    - Specific types of outliers and noise
    - True change labels are unavailable
  - 3+ component prediction system
    - Signal model (2<sup>nd</sup> order regression)
    - Outlier elimination
    - Change detection and training window selection
    - Backtracking
- Based on moving average and learnable thresholds
- 

Pechenizhkiy et al, 2009. SIGKDD Explorations 11(2), p. 109-116



# Stock Balancing Problem

Empty shelves

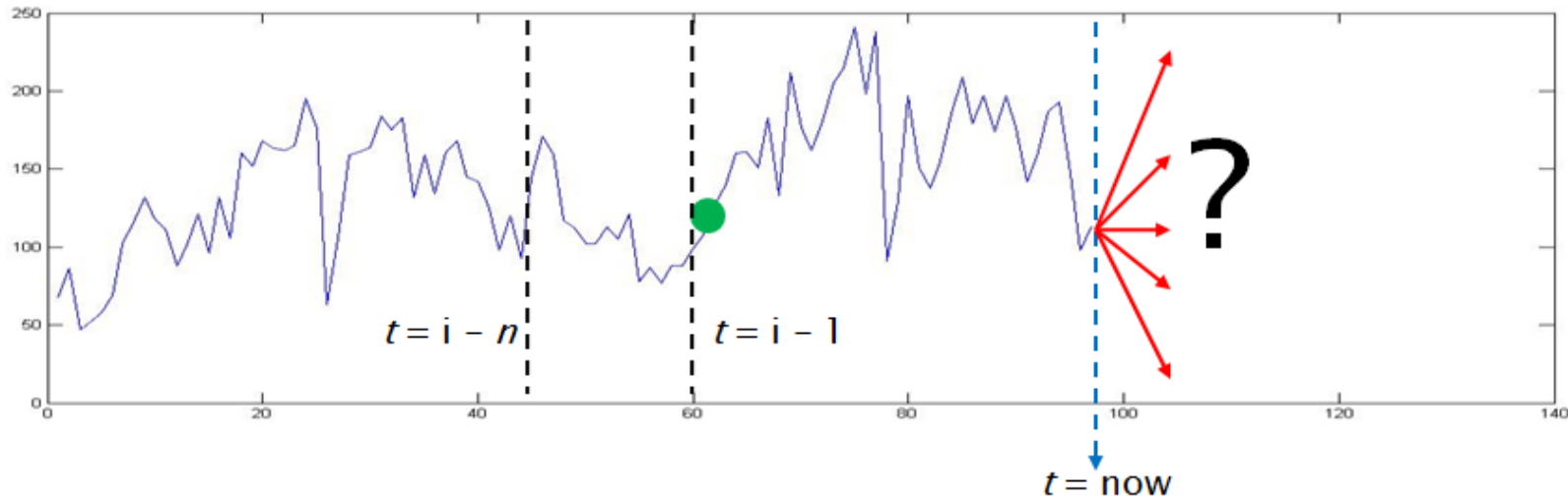


VS.



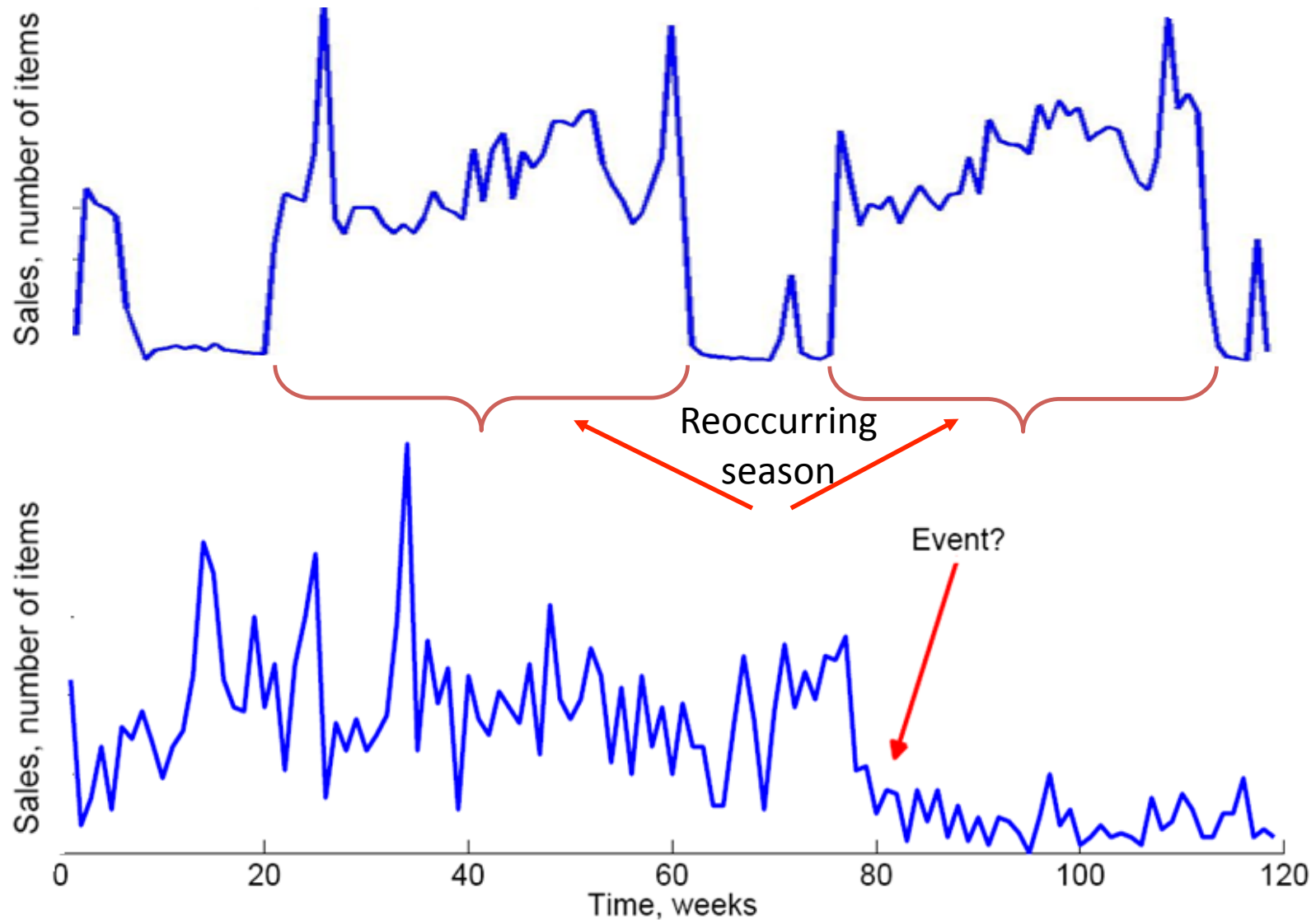
Perishable goods becoming obsolete

# Challenges in food sales prediction



| $t$                   | History                | Temp | Holiday | Promo |                                    |
|-----------------------|------------------------|------|---------|-------|------------------------------------|
| 1                     |                        |      |         |       |                                    |
| .                     |                        |      |         |       |                                    |
| <b><math>i</math></b> | $\{y(i-n) .. y(i-1)\}$ |      |         |       | Predict label of each new instance |
| .                     |                        |      |         |       |                                    |
| $n$                   |                        |      |         |       |                                    |

# Reoccurring and suddent drift in food sales

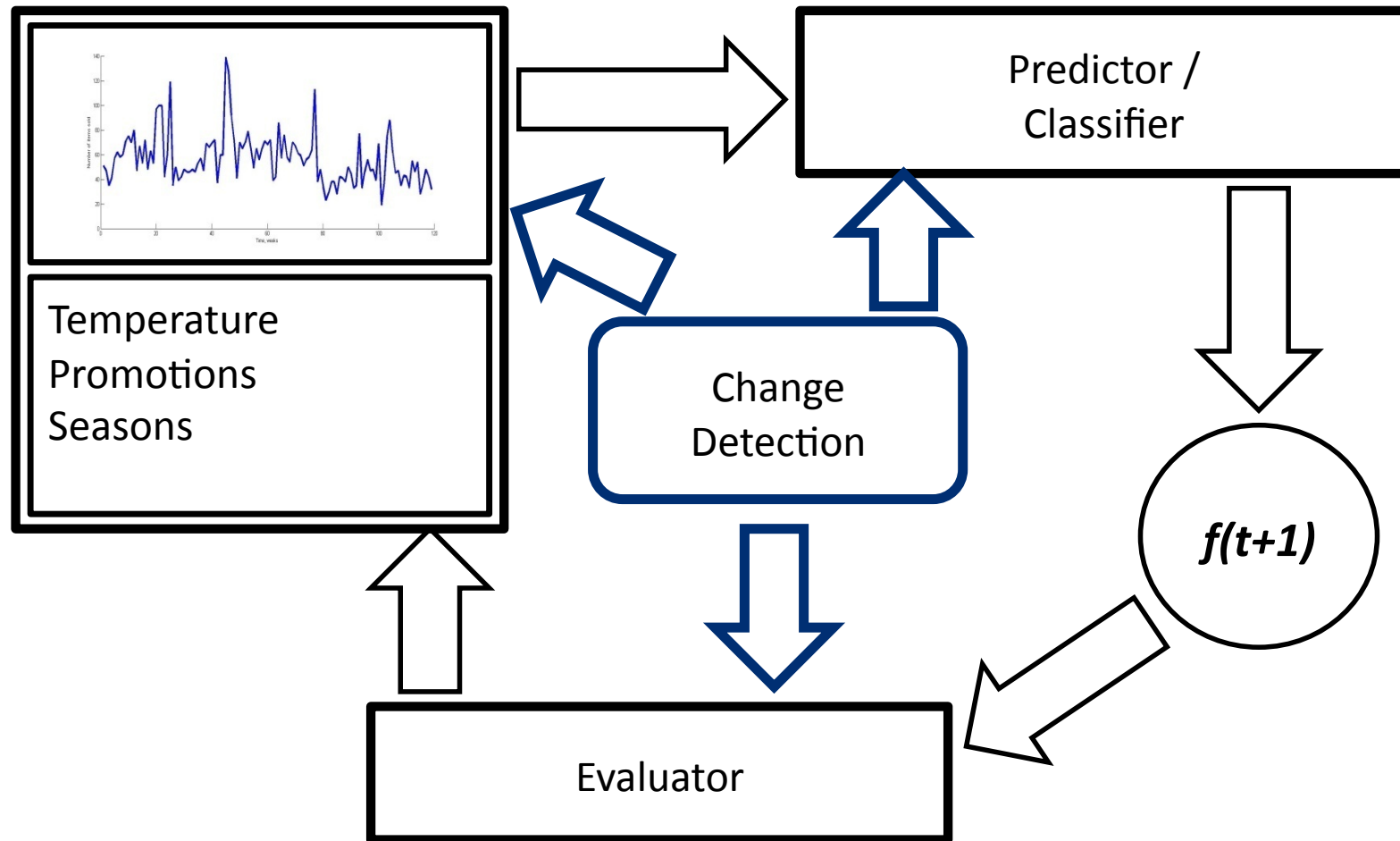


# Solution

- Adaptive learning approach
  - Added external data (weather, holidays)
  - Contextual (meta learning) approach
- Approach
  - Form contextual features (structural, shape, relational)
  - Identify product categories
  - Train a switch mechanism, which assigns the predictor, based on what context is observed
- Challenges
  - Short history, rapid and frequent changes
  - Discretization, formulating the labels

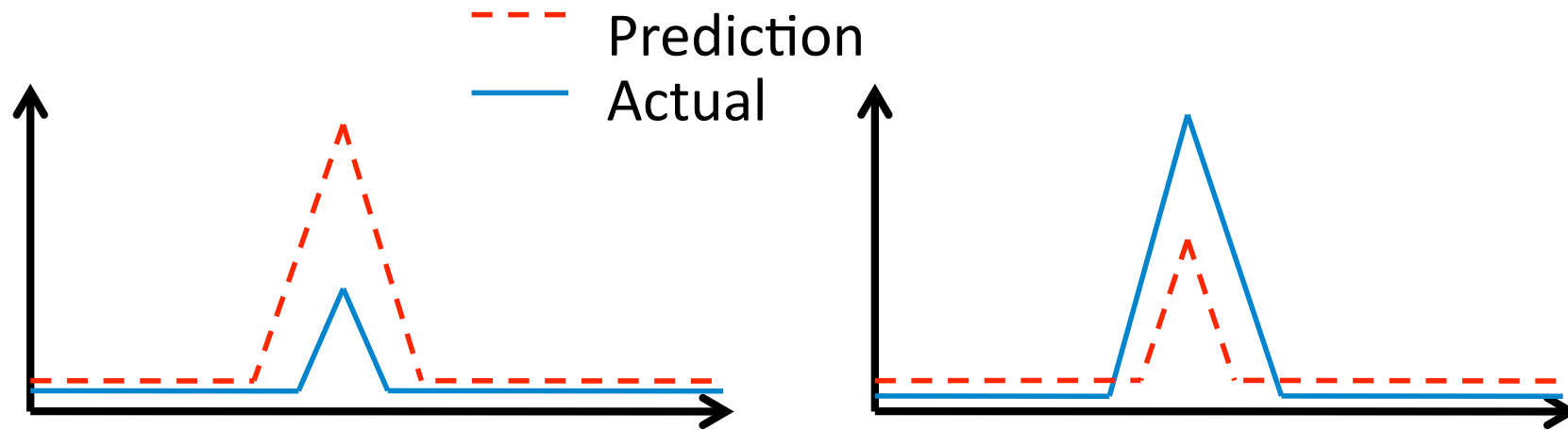
Žliobaitė et al, 2010. Beating the baseline prediction in food sales: How intelligent an intelligent predictor is?

# Solution





# Different Kinds of Errors



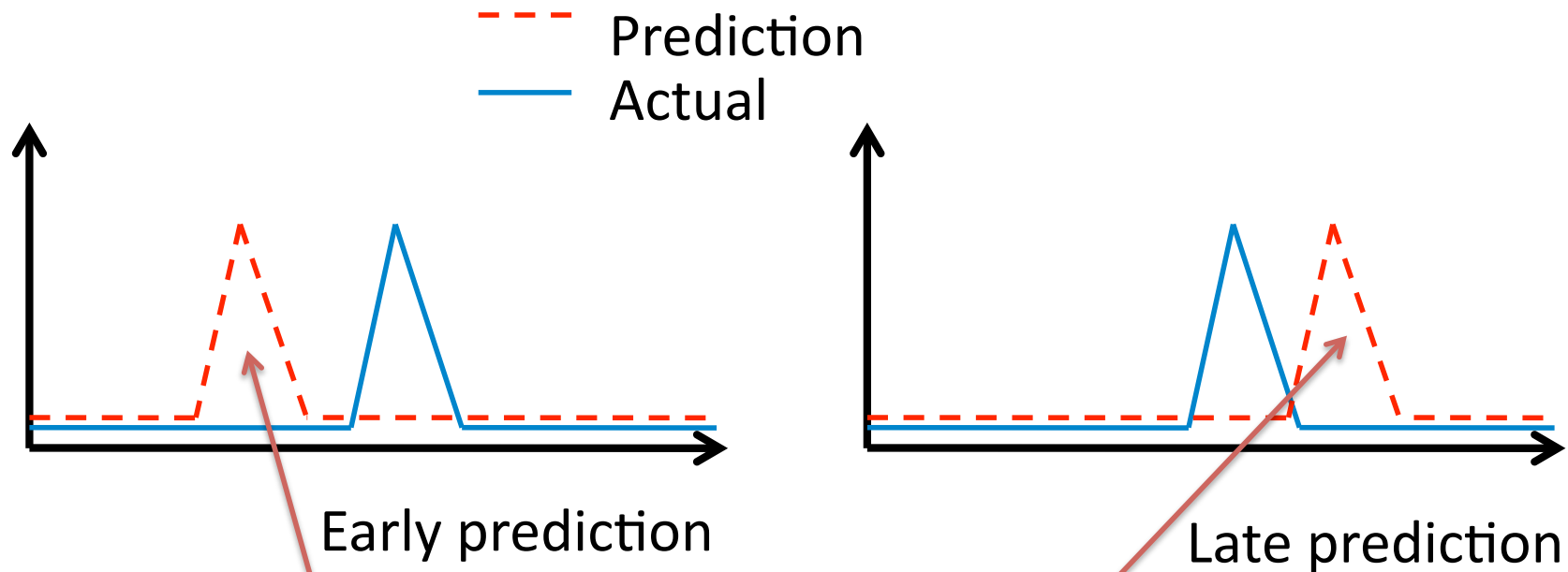
Over prediction

Over prediction might lead to overstocking and generate more costs

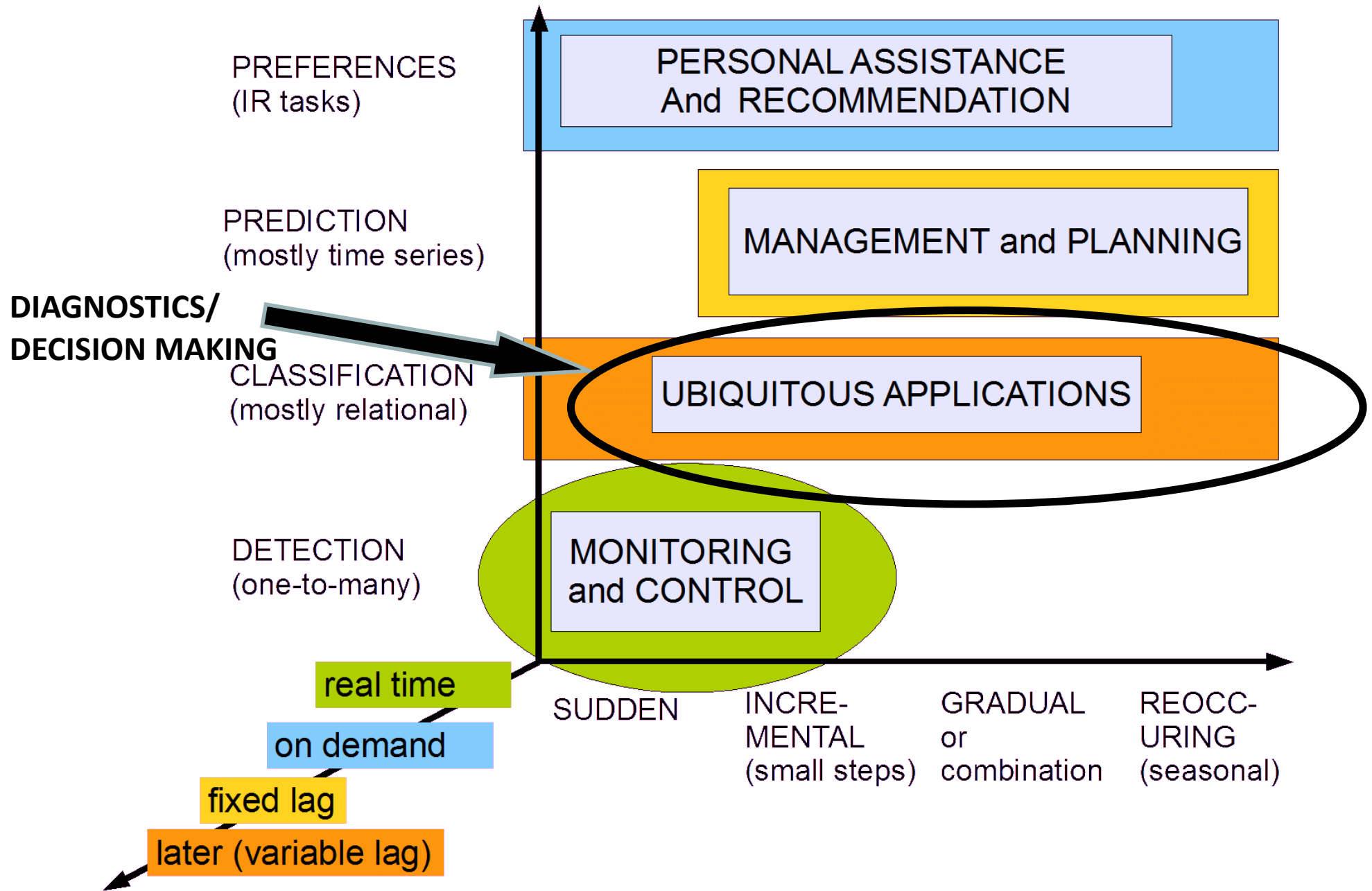
Under prediction

Under prediction might lead to understocking and generate more costs

# Different Kinds of Errors



Predicting high demand for Wednesday rather than Monday while it is in fact on Tuesday is usually worse (likely to generate more costs)



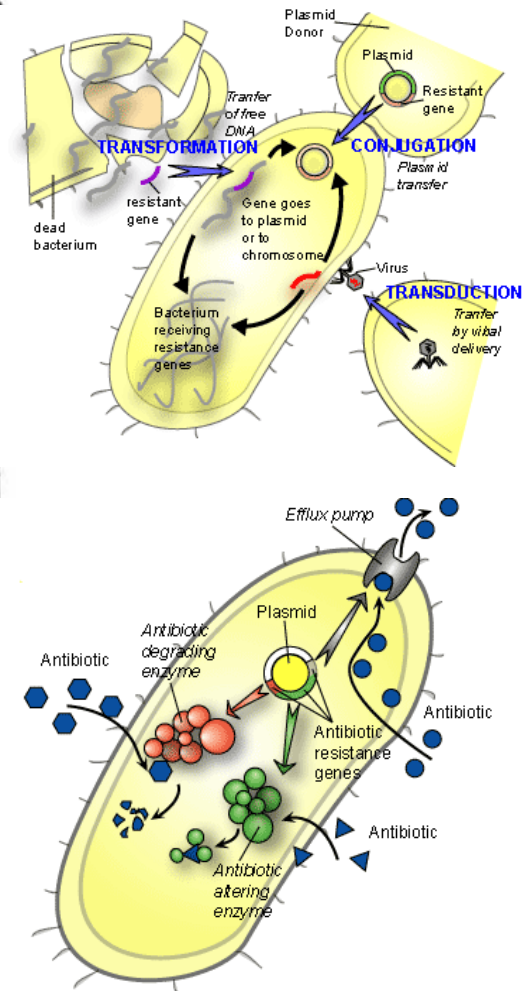
# Antibiotic Resistance Prediction

predict the sensitivity of a pathogen to an antibiotic based on data about the antibiotic, the isolated pathogen, and the demographic and clinical features of the patient.

| date      | sex | age | isNew | days_total | days_ICU | main_dept | pathogen data | antibiotic data | sensitivity |
|-----------|-----|-----|-------|------------|----------|-----------|---------------|-----------------|-------------|
| 22.1.2002 | m   | 25  | 1     | 171        | 81       | 9         | ...           | ...             | 3           |
| 22.1.2002 | m   | 25  | 1     | 171        | 81       | 9         | ...           | ...             | 3           |
| 22.1.2002 | m   | 25  | 1     | 171        | 81       | 9         | ...           | ...             | 3           |
| 22.1.2002 | m   | 25  | 1     | 171        | 81       | 9         | ...           | ...             | 3           |
| 28.1.2002 | f   | 61  | 0     | 261        | 52       | 3         | ...           | ...             | 3           |
| 28.1.2002 | f   | 61  | 0     | 261        | 52       | 3         | ...           | ...             | 3           |
| 28.1.2002 | f   | 61  | 0     | 261        | 52       | 3         | ...           | ...             | 3           |
| 28.1.2002 | f   | 61  | 0     | 261        | 52       | 3         | ...           | ...             | 1           |
| 28.1.2002 | f   | 61  | 0     | 261        | 52       | 3         | ...           | ...             | 1           |
| 28.1.2002 | m   | 25  | 1     | 171        | 81       | 9         | ...           | ...             | 3           |
| 28.1.2002 | m   | 25  | 1     | 171        | 81       | 9         | ...           | ...             | 3           |
| 30.1.2002 | m   | 25  | 1     | 171        | 81       | 9         | ...           | ...             | 3           |
| 8.2.2002  | m   | 30  | 0     | 209        | 209      | 9         | ...           | ...             | 3           |
| 8.2.2002  | m   | 30  | 0     | 209        | 209      | 9         | ...           | ...             | 1           |
| 8.2.2002  | m   | 30  | 0     | 209        | 209      | 9         | ...           | ...             | 1           |
| 11.2.2002 | f   | 0   | 0     | 18         | 0        | 2         | ...           | ...             | 1           |
| 11.2.2002 | f   | 0   | 0     | 18         | 0        | 2         | ...           | ...             | 1           |
| 11.2.2002 | f   | 0   | 0     | 18         | 0        | 2         | ...           | ...             | 1           |
| new data  | ... | ... | ...   | ...        | ...      | ...       | ...           | ...             | ?           |
| new data  | ... | ... | ...   | ...        | ...      | ...       | ...           | ...             | ?           |
| new data  | ... | ... | ...   | ...        | ...      | ...       | ...           | ...             | ?           |

(Tsymbal et al., 2008; Information Fusion 9(1))

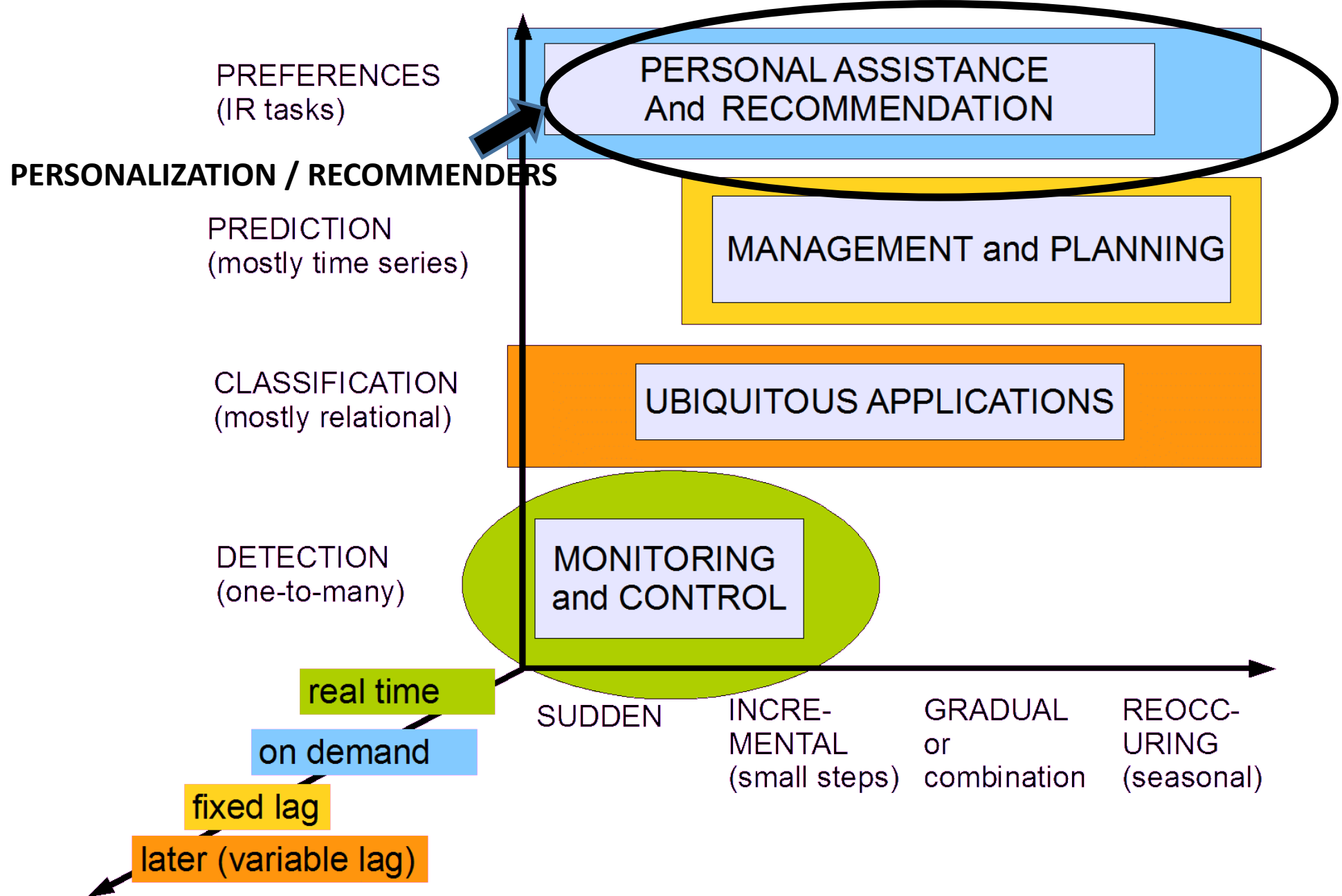
# How Antibiotic Resistance Happens



It was on a short-cut through the hospital kitchens that Albert was first approached by a member of the Antibiotic Resistance.

# Challenges and Solution

- Different ways how resistance may happen
- Different ways how resistance may spread
- Physical connections/hospital organization
- Local drift
- Solution:
  - Contextual approach and the instance level
  - Dynamic integration of classifiers



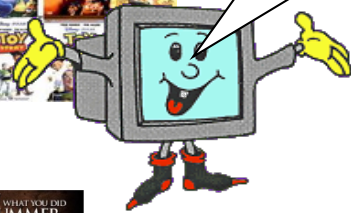
# Recommender Systems

## Lessons learnt from Netflix competition:

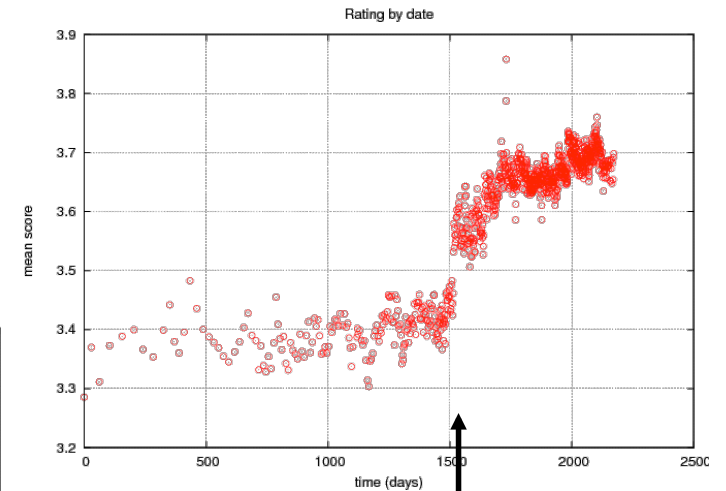
Temporal dynamics is important

Classical CD approaches may not work

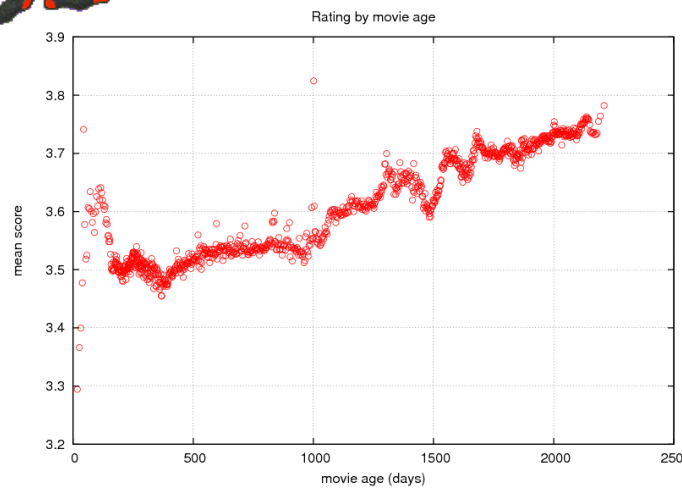
We Know What You Ought To Be Watching This Summer



(Koren, SIGKDD 2009)



Something happened in early 2004...



Are movies getting better with time?





# Multiple sources of temporal dynamics

- Both items and users are changing over time
  - Item-side effects:
    - Product perception and popularity are constantly changing
    - Seasonal patterns influence items' popularity
  - User-side effects:
    - Customers ever redefine their taste
    - Transient, short-term bias; anchoring
    - Drifting rating scale
    - Change of rater within household

# Block 4: Summary

- Working with real data and applications in which CD occurs and matters is fun and full of surprises
  - will always strengthen your results, and
  - may suggest you new research problems (driven by real challenges) to address
- We presented a reference framework for categorizing different (types of) applications
  - data/task/goal, expectations about changes, and operational settings dimensions
  - don't use it as a ground truth but as a guide helping to better understand the landscape of existing and foreseen problems and to better position your own problem/work

# Current and Foreseen Promising Directions

- Changing the focus: from blind adaptivity to
  - change modeling and description
  - recognizing and reusing similar situations from the past
- Application driven concept drift problems, like
  - label unavailability or lag
  - cost-benefit trade off of the model update
  - controlled adaptivity (due to adversaries)
  - lack of ground truth for training
- A reference framework and guidelines
  - for incorporating adaptivity in modeling
  - to be used in different application tasks

# Tutorial Summary

- Data patterns change over time,
  - prediction systems need to be adaptive
  - to maintain accuracy
- Four types of learning techniques
  - make different assumptions about the data and change,
  - may be appropriate in different situations
- Application tasks have different properties therefore
  - pose different challenges,
  - may require different handling techniques,
  - there is no “one fits all solution”
- Design of adaptive prediction systems shall be tightly connected with application task at hand

# Tutorial Summary

- Current state-of-the-art in handling CD
- Categorization of approaches and categorization of applications
- Popular techniques
- Lessons learnt from case studies
- MOA software framework – contribute your techniques
- Initiative for the repository of available benchmarks for CD research – contribute your datasets
- Promising directions for further research from the applications perspective

# HaCDAIS@IEEE ICDM 2011

- 2nd Int. Workshop on Handling Concept Drift in Adaptive Information Systems
  - Vancouver, Canada, December 10th, 2011.
  - in conjunction with the 11th IEEE International Conference on Data Mining (IEEE ICDM 2011).
  - Welcome to participate
  - Welcome to contribute:
    - July 23, 2011 - Submissions due
- More info:
  - <http://wwwis.win.tue.nl/hacdais2011/>

# Data Streams @ ACM SAC 2012

- Data Streams Track
  - ACM Symposium on Applied Computing
- Trento University, Italy, March 20-23, 2012.
  - Paper Submission: 31 August, 2011
- Welcome to contribute!
  - <http://www.cs.waikato.ac.nz/~abifet/SAC2012/>

# Bibliography – General Background

- *Detection of Abrupt Changes - Theory and Application*, M. Basseville, I. Nikiforov, Prentice-Hall, Inc. 1993.
- *Statistical Quality Control*, E. Grant, R. Leavenworth, McGraw-Hill, 1996.
- *Continuous Inspection Scheme*, E. Page. *Biometrika* 41 1954
- *Learning in the Presence of Concept Drift and Hidden Contexts*, G. Widmer, M. Kubat: *Machine Learning* 23(1): 69-101 (1996)
- *Learning drifting concepts: Example selection vs. example weighting*, R. Klinkenberg, IDA 2004
- *Learning with Drift Detection*; J. Gama, P. Medas, G. Castillo, P. Rodrigues; SBIA 2004, Springer.
- *The problem of concept drift: Definitions and related work*, Tsymbal, A. Technical Report. Dept. Computer Science, Trinity College, Ireland, 2004.
- *Adaptive Learning and Mining for Data Streams and Frequent Patterns*, Albert Bifet, PhD Thesis, 2009
- *Adaptive Training Set Formation*, Indre Žliobaitė, PhD Thesis, Vilnius University, Lithuania, 2010.



# Bibliography - Approaches

- *Mining Time-Changing Data Streams*, G. Hulten, L. Spencer, P. Domingos, ACM SIGKDD, 2001.
- *Dynamic Weighted Majority: A New Ensemble Method for Tracking Concept Drift*, by J. Kolter, M. Maloof, ICDM 2003.
- *Mining Concept Drifting Data Streams using Ensemble Classifiers*, by H.Wang, Wei Fan, P. Yu, J. Han, ACM SIGKDD 2003.
- *Decision Trees for Mining Data Streams*; J. Gama, R. Fernandes, R.Rocha. *Intelligent Data Analysis* 10(1):23-45 (2006)
- *OLINDDA: A cluster-based approach for detecting novelty and concept drift in data streams*. Spinosa, E.J., Carvalho, A., and Gama, J. 22ndACM SAC 2007: ACM Press.
- *An Ensemble of Classifiers for Coping with Recurring Contexts in Data Streams*, Katakis, I., Tsoumakas, G., and Vlahavas, I.. in 18th ECAI. 2008, IOS
- ...

# Bibliography - Applications

- *Reference Framework for Handling Concept Drift: An Application Perspective.* Žliobaitė, I. and Pechenizkiy, M. Technical report, Eindhoven University of Technology, 2010
- *Sentiment knowledge discovery in Twitter streaming data.* Albert Bifet and Eibe Frank. Proc 13th International Conference on Discovery Science, Canberra, Australia, 2010.
- *Collaborative Filtering with Temporal Dynamics* Yehuda Koren, KDD 2009, ACM, 2009
- *MOA: Massive online analysis, a framework for stream classification and clustering.* Albert Bifet, Geoff Holmes, Bernhard Pfahringer, Philipp Kranen, Hardy Kremer, Timm Jansen, Thomas Seidl, in Pechenizkiy, M. & Žliobaitė, I. (eds), HaCDAIS Workshop ECML-PKDD 2010, 2010
- *Online Mass Flow Prediction in CFB Boilers with Explicit Detection of Sudden Concept Drift.* Pechenizkiy, M., Bakker, J., Žliobaitė, I., Ivannikov, A., Karkkainen, T. SIGKDD Explorations 11(2), p. 109-116, 2009.
- *Dynamic Integration of Classifiers for Handling Concept Drift,* Tsymbal, A., Pechenizkiy, M., Cunningham, P. & Puuronen, S. Information Fusion, Special Issue on Applications of Ensemble Methods, 9(1), pp. 56-68, 2008.

# Acknowledgements

- Thanks to:
  - Raquel Sebastião, Pedro Rodrigues, Jorn Bakker
  - FCT financial support of LIAAD, and Project Kdus
  - Part of the research leading to these results has received funding from the European Commission within the Marie Curie Industry and Academia Partnerships and Pathways (IAPP) programme under grant agreement no. 251617.
  - NWO HaCDAIS Project

