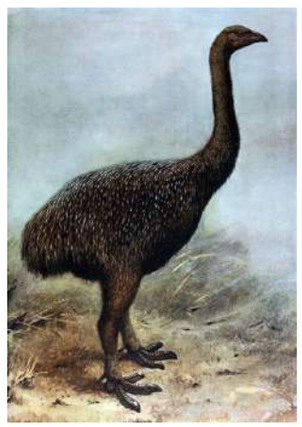
IncrLearn Workshop
ICDM 2021

Learning for Delayed Partially Labelled Data Streams

Heitor Murilo Gomes
University of Waikato
AI Institute

http://heitorgomes.com



Heitor Murilo Gomes

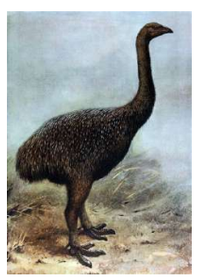
PhD in CS (Brazil - France) - 2017

Before: Postdoc at Télécom ParisTech (Paris, France)

Current: Senior Research Fellow at the University of Waikato (Hamilton, New Zealand)

Research: Machine learning (ML), Data streams, Ensemble Learning, Semi-Supervised Learning, Distributed ML

Head of the MOA lab and Co-director of the AI Institute



http://www.heitorgomes.com

Te Ipu o te Mahara
Artificial Intelligence
Institute
THE UNIVERSITY OF WAIKATO

Learning for Delayed Partially Labelled Data Streams

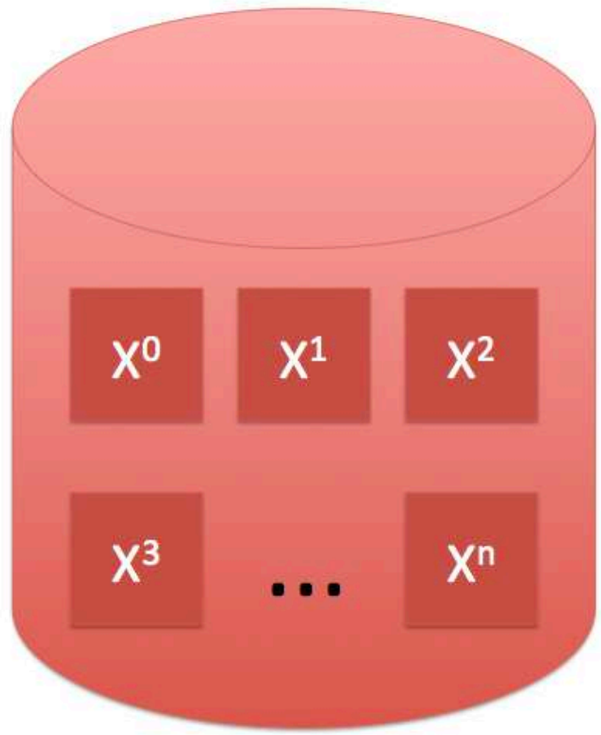
1. Learning from data streams
2. Delayed and partially labelled
3. SSL Strategies for DS
4. Delayed and SSL Drift Detection
5. Conclusions and perspectives

1. Learning from Data Streams

Batch data

Well defined training phase

Random access to instances



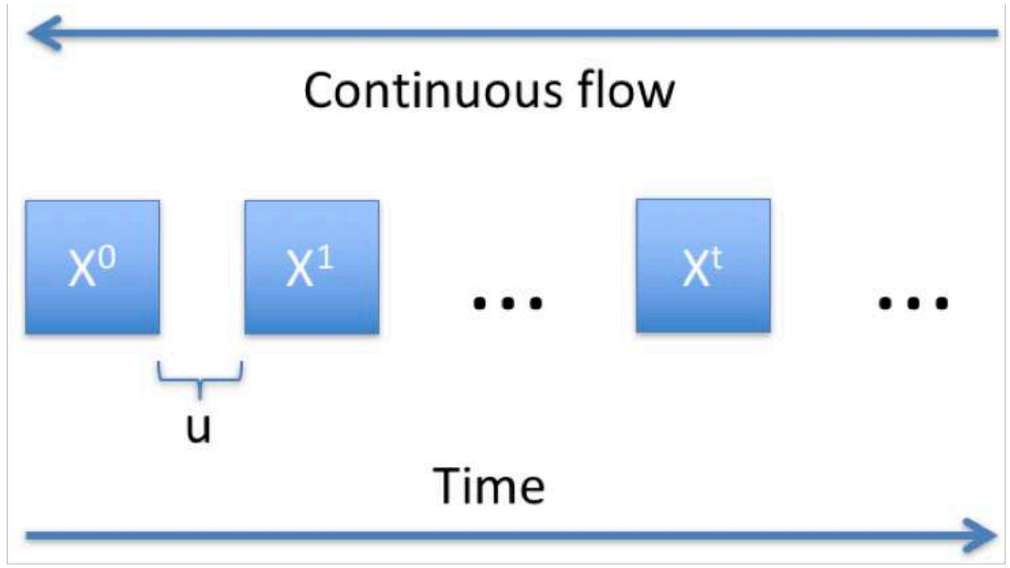
Challenges:
missing data, noise, imbalance, high dimensionality, ...

Learning from Data Streams

Streaming Data

Sequential access only

Strict time/memory requirements



Non-stationary data distribution

Challenges: inherit those from batch + concept drifts, feature evolution, ...

Offline (Batch) vs. Online (Streaming)

Offline (Batch) data

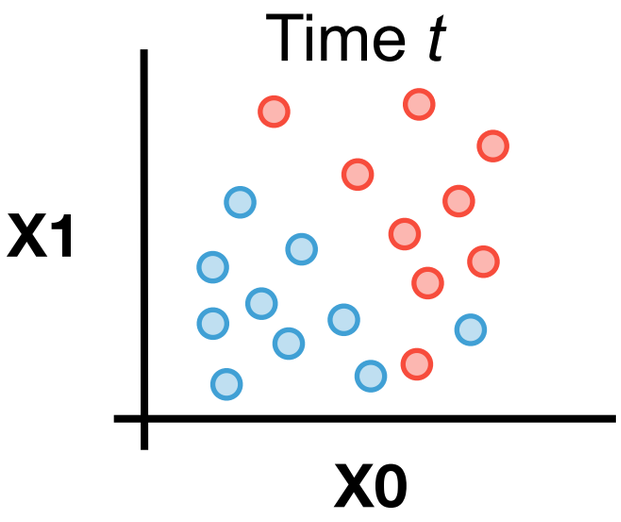
The output is a **trained model**

Online (Streaming) data

...

The output is a **trainable model**

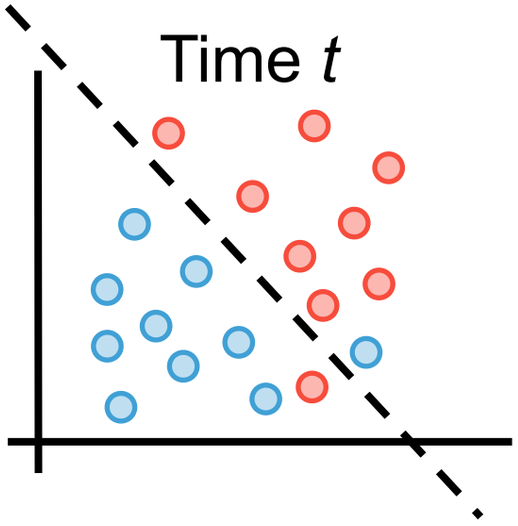
Concept drift



Assume this simple classification problem

- Two classes
- Two features (X0 and X1)

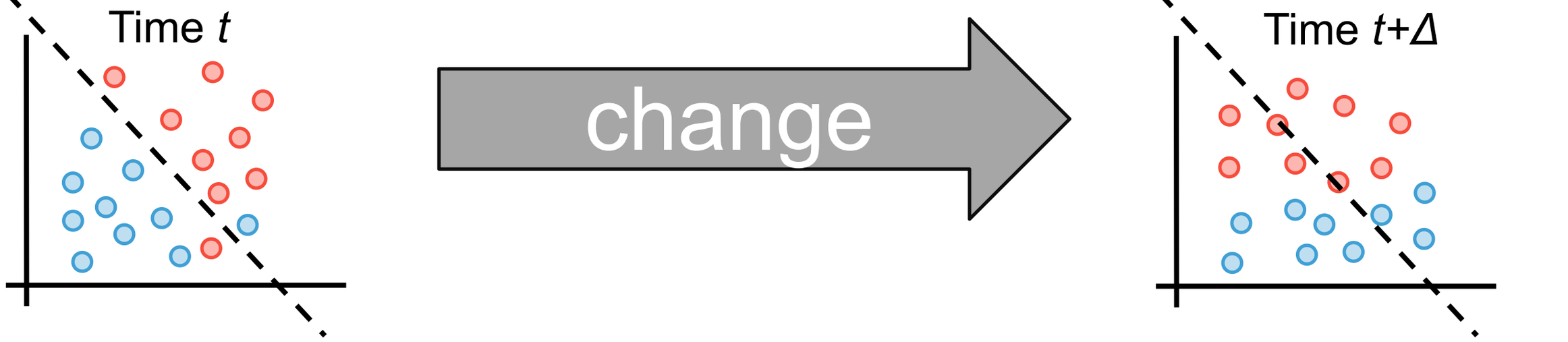
Concept drift



We can build a very simple
linear model to separate the two classes!

Concept drift

After a while, the data distribution changes…



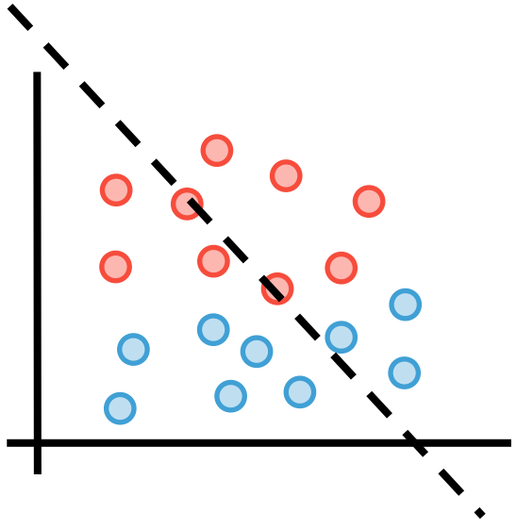
We had an accurate model…

Not anymore...

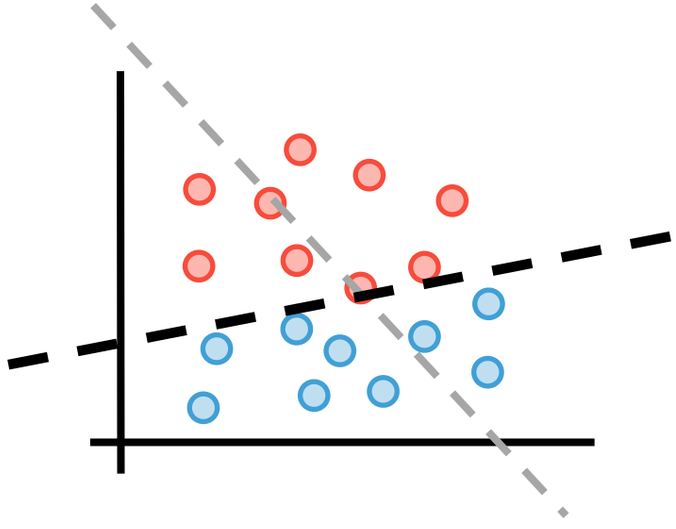
WHAT NOW?

Concept drift

We can detected and adapt!



change



Underperforming model…

An updated model!

Handling concept drift

Concept drift. The data distribution is unknown and it may change

What can we do about **concept drift**?

- **Detect** it

- **Recover** from it

Handling concept drift

Detect

- CUSUM, DDM, EDDM, ADWIN, …

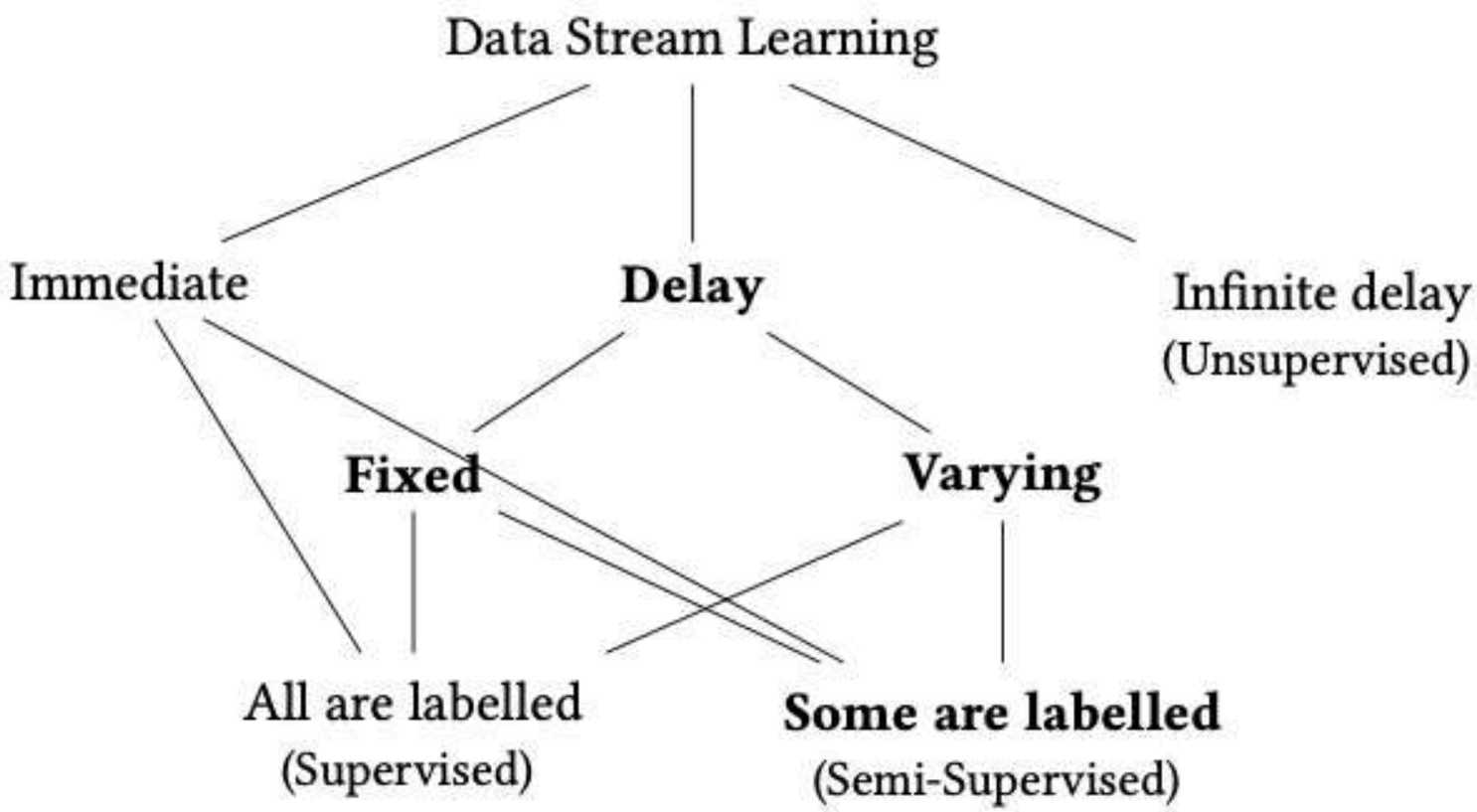
Recover (adapt and retrain)

- Reactive (signalled resets) - e.g. Adaptive Random Forest

- Active (periodical resets) - e.g. Sliding window kNN

2. Delayed and Partially Labelled Data Streams

Learning from data streams according to **labels arrival time**.



Highlighted in **bold** the dimensions associated with **delayed partially** labelled data streams.

Problem setting

(i) Immediate and fully labelled

The verification latency between x and y corresponds exactly one time unit

(ii) Delayed and fully labelled

There is a delay between x and y limited by the finite range

(iii) Immediate and partially labeled

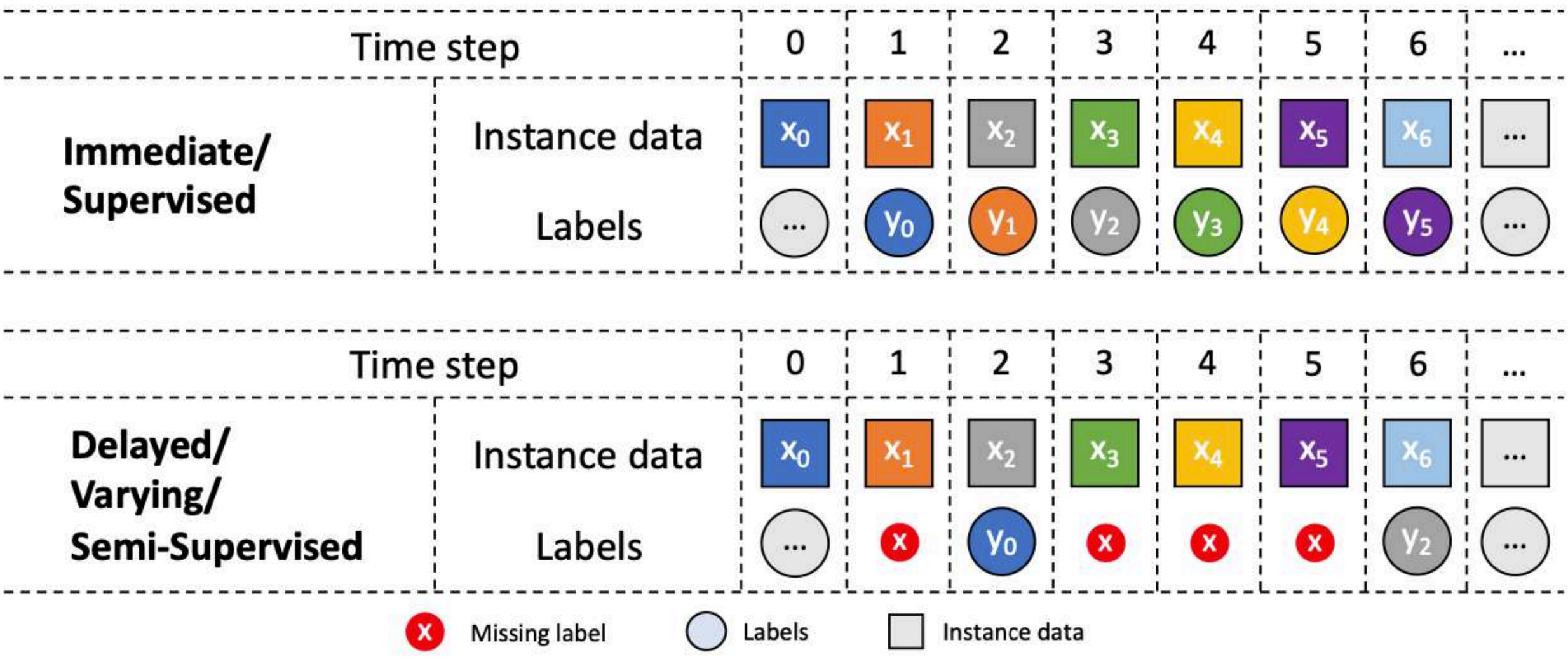
Relax the constraint that every x has a corresponding entry on y

(iv) Delayed and partially labelled

Similar to (iii), it relaxes the constraints from (ii)

A Survey on Semi-Supervised Learning for Delayed Partially Labelled Data Streams. Gomes, H. M., Grzenda, M., Mello, R., Read, J., Nguyen, M. H. L., & Bifet, A. (2021). *arXiv preprint arXiv:2106.09170*.

Problem setting: Example



Related problems

Active learning

- Overcome the labelling bottleneck by querying the label of selected unlabelled instances to an oracle

- The instances to be labelled are the most uncertain and can bring the highest value to the learning process.

- The active learner aims to achieve **high accuracy using as few labelled instances** as possible, thereby minimizing the cost of collecting labelled data.

- **Active learning may not be applicable for many streaming scenarios.**

(1) The oracle's response time may be too slow, as it often relies on a human expert.
(2) If a concept drift occurs, the instances selected to be labelled may be outdated. The latter issue can be amended by using active learning strategies that take drift into account [1]

[1] **Active learning with drifting streaming data.** Žliobaitė, Indrė, et al. *IEEE transactions on neural networks and learning systems* 25.1 (2013): 27-39.

Related problems

Transductive learning. The unlabelled test data set contains the whole of instances to be predicted, thus instead of producing a general model for predicting any future instance, the output is the predictions. This is a "closed world" assumption, where a successful solution is one where the algorithm can approximate the true labels of the instances solely for the finite test data set.

Initially Labelled Streaming Environment. Labelled data may only be available at the beginning of the learning process. Therefore, a supervised learning algorithm can be trained with the initial data, and another unsupervised mechanism used to update the model during execution.

Concept evolution. Some instances are not only unlabelled but belong to a class that has not yet been identified.

3. SSL Strategies for DS

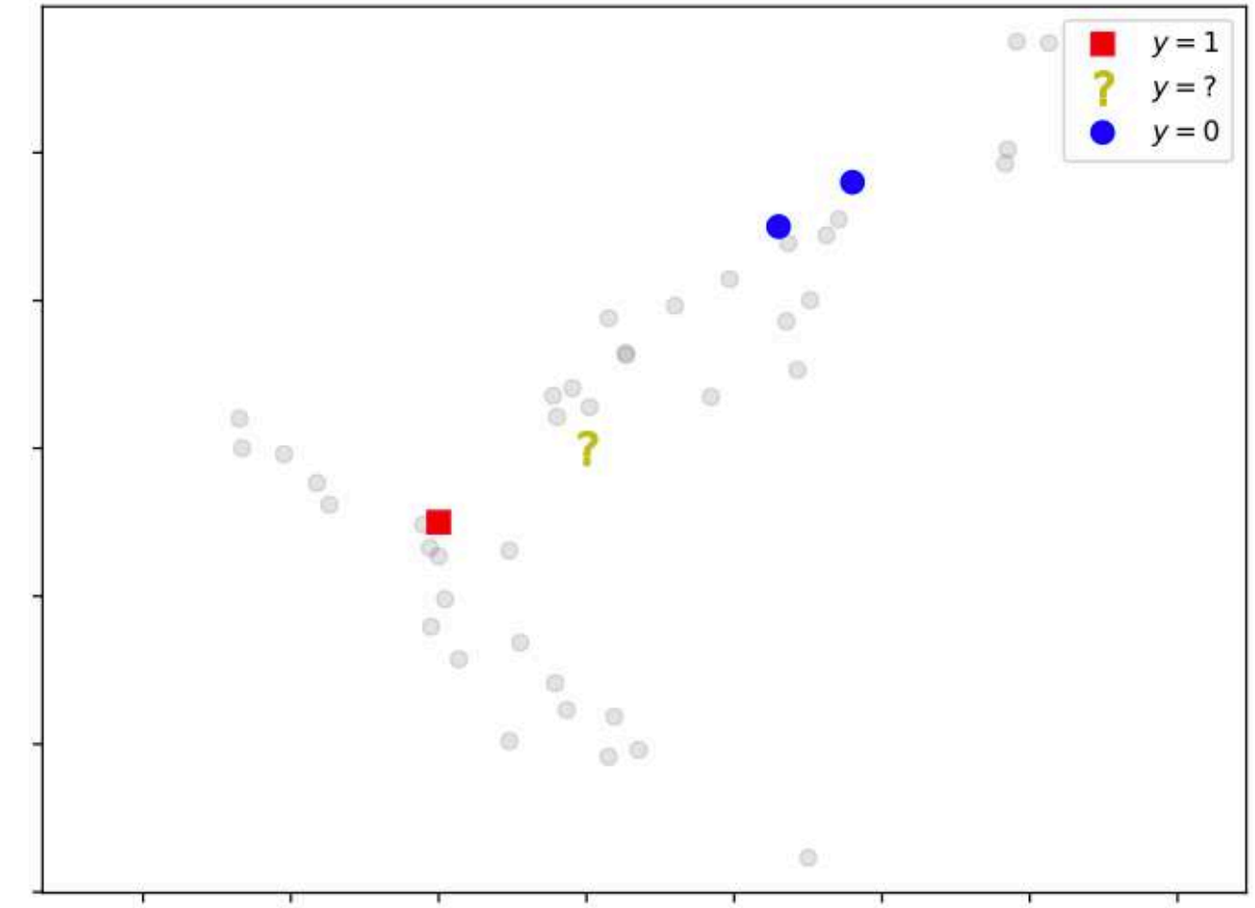
- SSL is a paradigm of learning that exploit unlabelled data to leverage models trained with labelled data

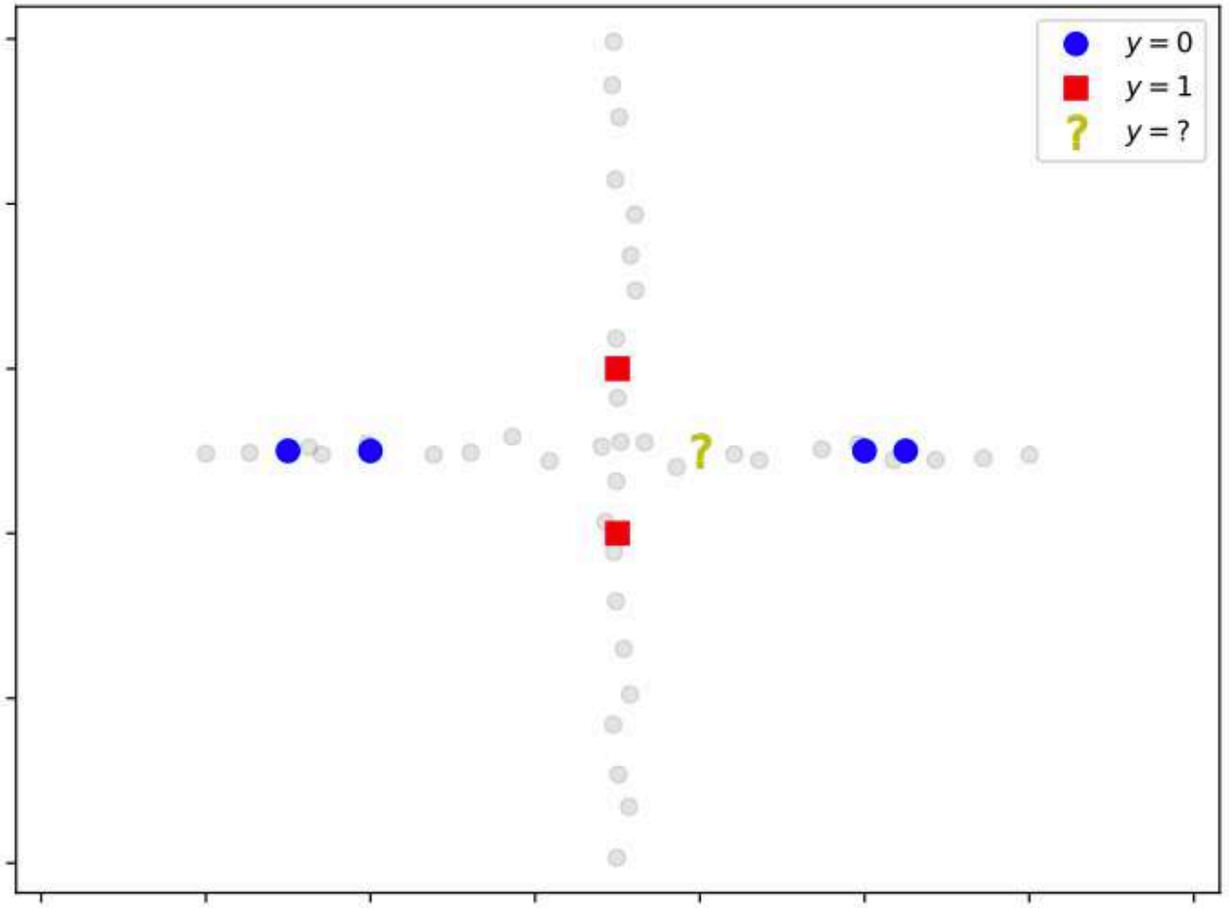
- The caveat is that SSL methods make strong assumptions about the data or the model

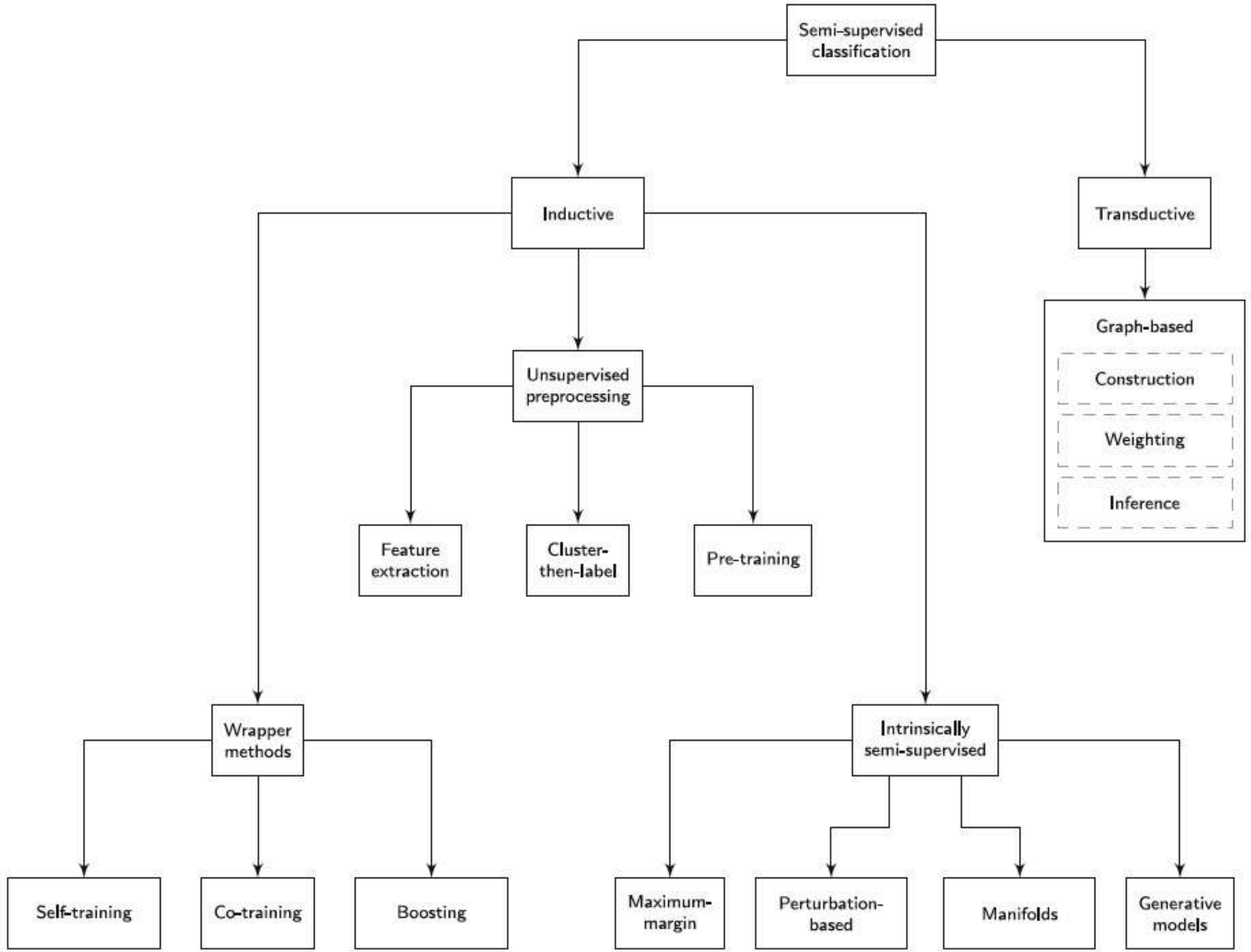
A survey on semi-supervised learning. Van Engelen, Jesper E., and Holger H. Hoos. *Machine Learning* 109.2 (2020): 373-440.

Example

Common assumptions: underlying density across points belongs to a single class, or a common manifold underlying the points of each class







Visualization of the semi-supervised classification taxonomy from [1]

[1] **A survey on semi-supervised learning.** Van Engelen, Jesper E., and Holger H. Hoos. *Machine Learning* 109.2 (2020): 373-440.

Cluster-then-label

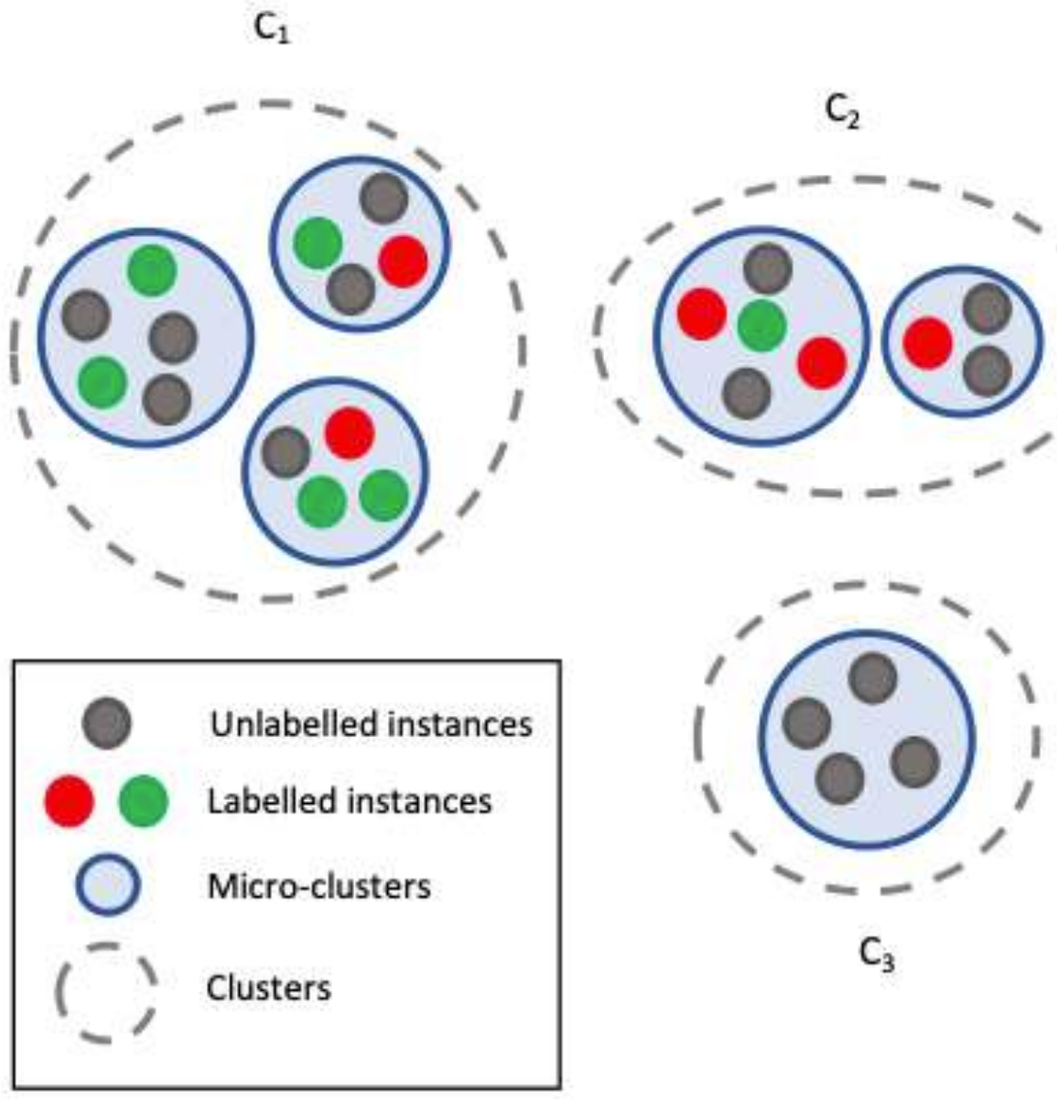
- Assumes that instances belonging to the same cluster may share the same label.

- Applying classic clustering algorithms, such as k-means, to streaming data is challenging as such algorithms repeatedly iterate over the data.

- The majority of the stream clustering methods incrementally update micro-clusters (summarised representations of the input data).

- The actual clustering algorithm is only occasionally executed in an offline step using the micro-clusters as input.

Cluster-then-label



An example of clusters, micro-clusters and instances. C1 and C2 clusters contain a majority of instances belonging to the green and red class labels, respectively. C3 has no labelled instances; thus no inference about the label of new instances assigned to it can be made.

Self-training

- Let a classifier learn from its previous mistakes and try to reinforce itself.

- Self-training acts as a wrapper algorithm that takes any arbitrary classifier.

- If we have an existing, fully-supervised learner that is complicated and hard to modify, self-training is an approach worth considering.

Self-training

- In [1], the authors proposed a self-training learner designed to receive as input either a single instance or a batch of instances at a time.

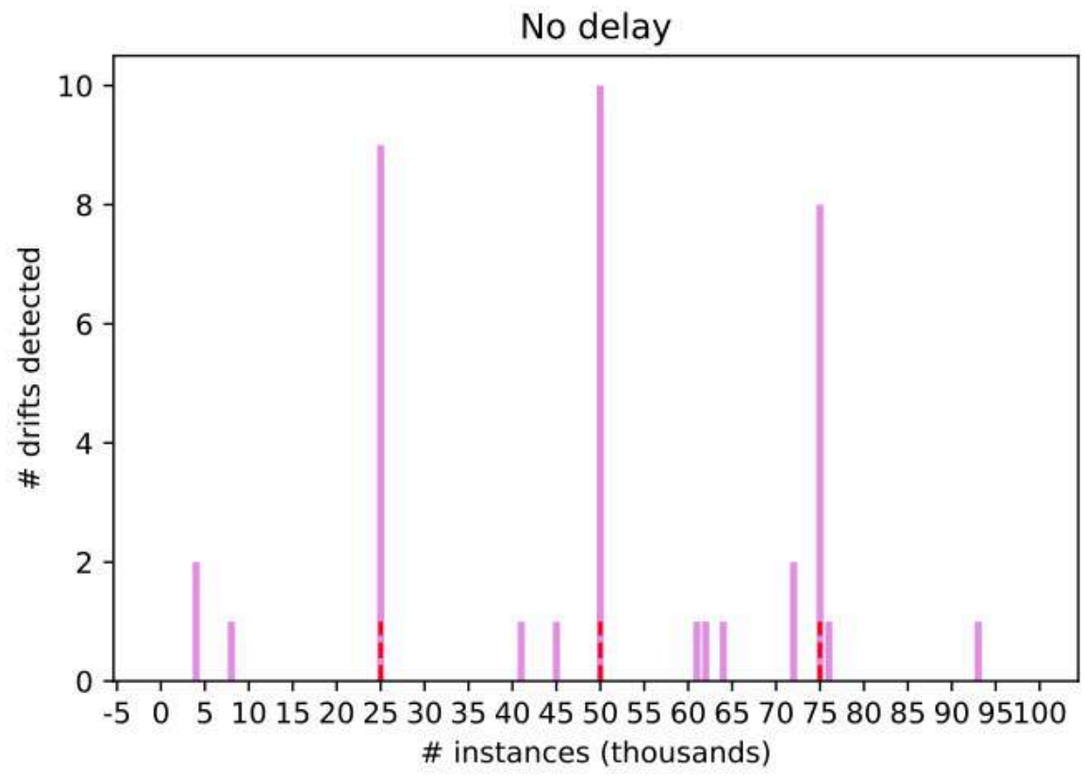
- The **confidence threshold** that determines whether instances are used for self-training could be fixed or adaptive concerning the average confidence scores observed in a window.

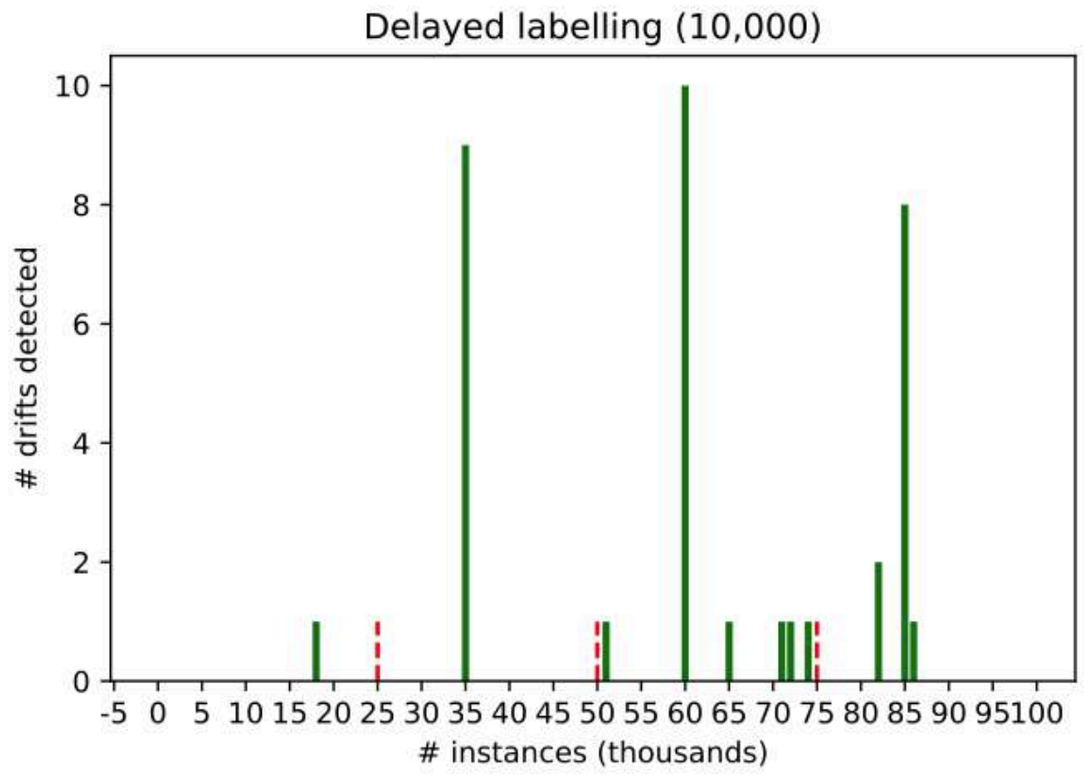
- Challenge: In batch data, the self-training process is iterative

[1] **Semi-supervised Learning over Streaming Data using MOA.** Le Nguyen, Minh Huong, et al. *IEEE International Conference on Big Data (Big Data)*. 2019.

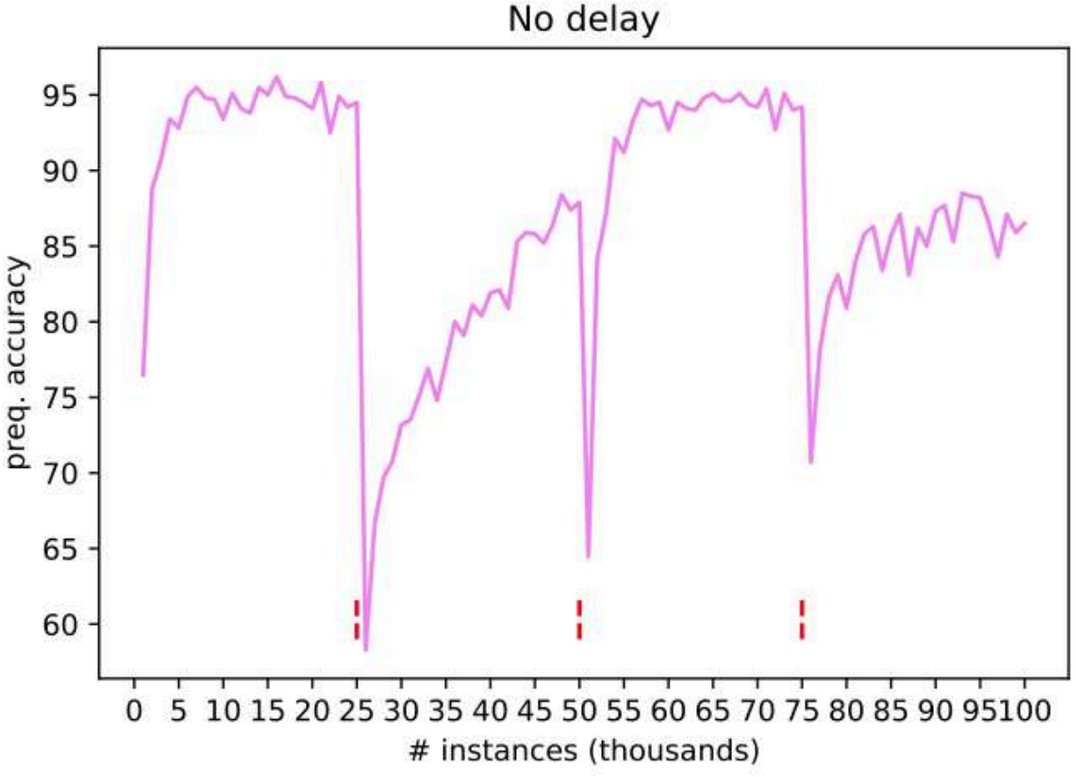
4. Delayed and SSL Drift Detection

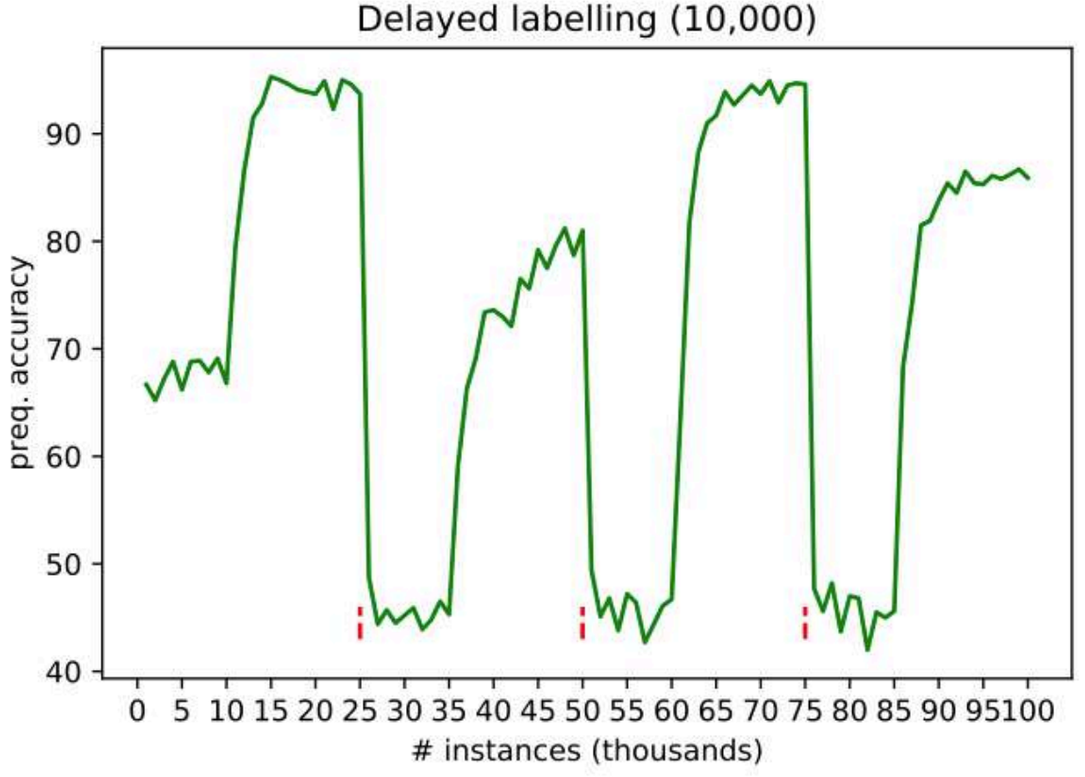
- Impact of delayed labelling on reactive strategies
- Conditions to detect concept drift in the absence of labels
- Strategies for detecting concept drift in such setting





Drifts detected by a 10 learner SRP model using ADWIN on AGRAWAL with and without labelling delay. Red dotted vertical lines indicate the location of concept drifts.





Accuracy by a 10 learner SRP model using ADWIN on AGRAWAL with and without labelling delay.

When concept drift detection is observable (and important)

- Indre Žliobaite [1] presented an analytical view of the conditions that must be met to allow concept drift detection in a delayed labelled setting.
- These conditions depends on characteristics of the drift, i.e. how the change affected the input probability P (X) and the posterior probabilities P (wi |X) of the class labels wi.

		$P(w_i\|X)$	
		change	**no change**
$P(X)$	**change**	(1) important & observable	(2) observable
	no change	(3) important & unobservable	(4) no drift

Change with delayed labeling: When is it detectable? Žliobaite, Indre. *2010 IEEE International Conference on Data Mining Workshops*.

When concept drift detection is observable (and important)

- One interesting connection between SSL and unsupervised drift detection is that if the underlying marginal data distribution P(X) over the input does not contain information about P(y|X) or indicates changes on P (y|X), then it is impossible to exploit unlabelled data to improve a supervised learner (SSL) or detect a change

		$P(w_i\|X)$	
		change	**no change**
$P(X)$	**change**	(1) important & observable	(2) observable
	no change	(3) important & unobservable	(4) no drift

Change with delayed labeling: When is it detectable? Žliobaite, Indre. *2010 IEEE International Conference on Data Mining Workshops*.

Teacher-student approach

- Cerqueira et al. [1] presents an unsupervised drift detector based on a teacher-student approach
- A predictive model (teacher) is built using an initial batch of labelled training data. The teacher's predictions are used as class labels to train a surrogate model (student), which will learn to mimic the teacher.
- A drift detection algorithm is used to identify variations in the mimicking error of the student.
- **Hypothesis:** if the mimicking error increases, then it means that a concept drift has occurred.

[1] **Unsupervised concept drift detection using a student–teacher approach.** Cerqueira, Vitor, et al. *International Conference on Discovery Science*. Springer, Cham, 2020.

Conclusions and perspectives

Evaluation process.

- Evaluating algorithms in a streaming setting is a challenging task
- The fair evaluation of delayed and partially labelled problems is even more intricate
- Some important aspects concern:
  - Document all the assumptions and limitations of the problem setting
  - Compare against “strong” supervised models trained only on the labelled data [1]

[1] **Realistic evaluation of deep semi-supervised learning algorithms.** Oliver, Avital, et al. *arXiv preprint arXiv:1804.09170* (2018).

Conclusions and perspectives

- Learning from data streams is becoming more relevant
- There is not enough labelled data (in most cases) available
- Often active learning or similar approaches are not applicable
- The combination of reliable SSL methods and unsupervised (or SSL) drift detection procedures seems like a reasonable path to explore this problem setting

[1] **Realistic evaluation of deep semi-supervised learning algorithms.** Oliver, Avital, et al. *arXiv preprint arXiv:1804.09170* (2018).

Thank you!

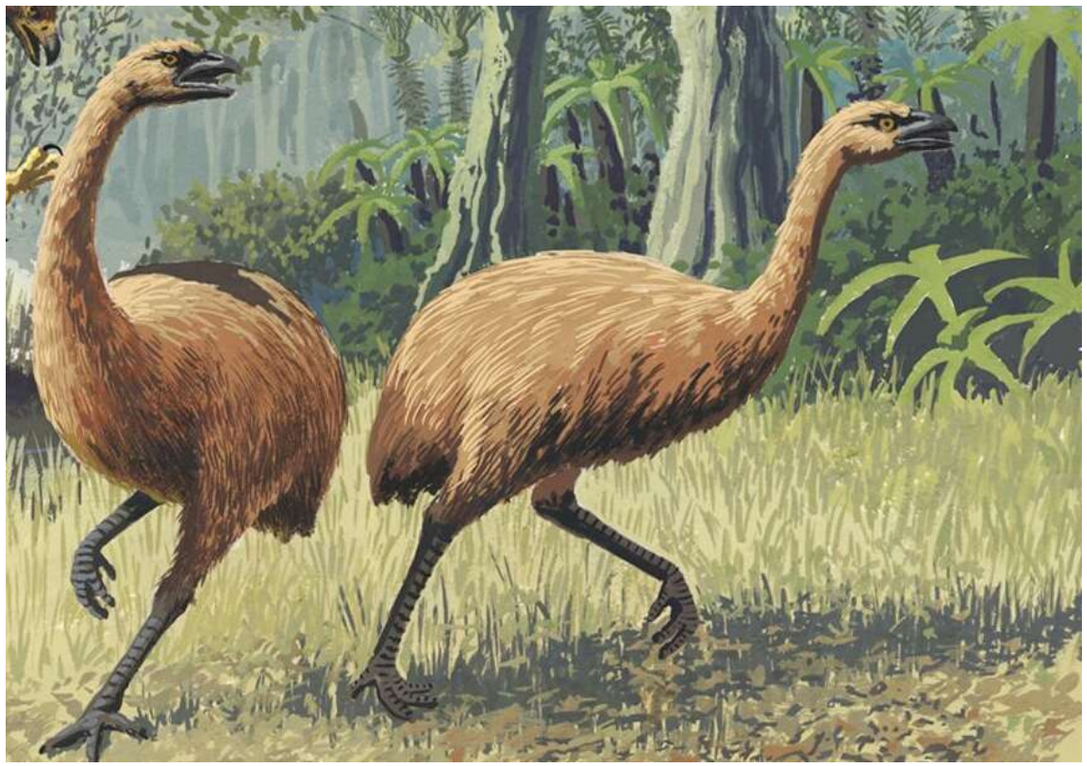


Image: John Megahan/PLOS Biology.

http://www.heitorgomes.com