Machine Learning for Streaming Data

IJCAI Tutorial 2024

Heitor Murilo Gomes[1*], Nuwan Gunasekara[2]
Albert Bifet[2,3], Bernhard Pfahringer[2]

* Corresponding author: heitor.gomes@vuw.ac.nz

https://nuwangunasekara.github.io/ijcai2024/

VICTORIA UNIVERSITY OF WELLINGTON
TE HERENGA WAKA

THE UNIVERSITY OF WAIKATO
Te Whare Wānanga o Waikato
KO TE TANGATA

TAIAO
Time-Evolving Data Science and Artificial Intelligence for Advanced Open Environmental Science

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

[1] Victoria University of Wellington, New Zealand, [2] University of Waikato, New Zealand,
[3] TELECOM Paris, LCTI, France.

Advanced topics

Partially and delayed labeled data

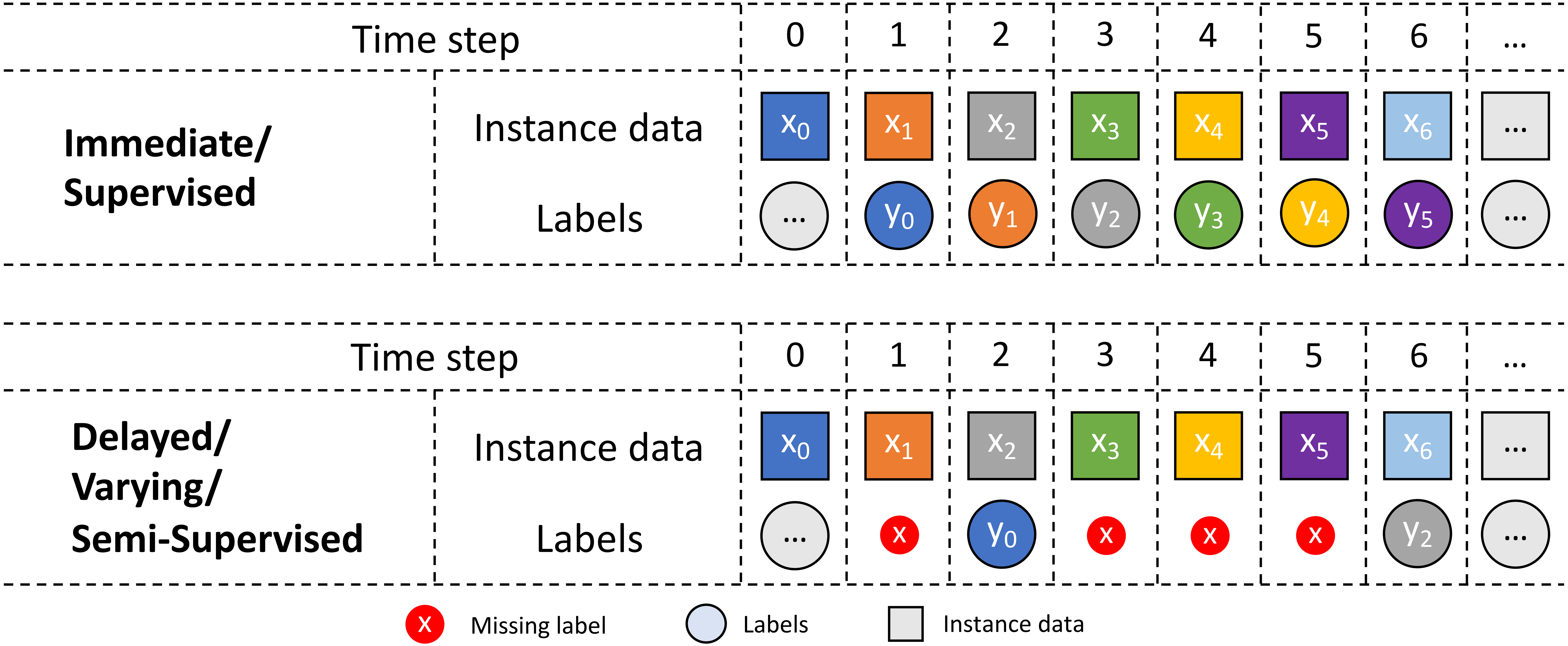
Partially labeled: Not all instances from the stream will be labeled

Delayed labeled: The label may take a while to arrive

Immediately labeled*: The class label arrives before the next input data to be predicted

Data streams are often associated with large amounts of data. However, most of this data is unlabelled

Partially and delayed labeled data



Time step		0	1	2	3	4	5	6	...
Immediate/ Supervised	Instance data	x_0	x_1	x_2	x_3	x_4	x_5	x_6	...
	Labels	...	y_0	y_1	y_2	y_3	y_4	y_5	...

Time step		0	1	2	3	4	5	6	...
Delayed/ Varying/ Semi-Supervised	Instance data	x_0	x_1	x_2	x_3	x_4	x_5	x_6	...
	Labels	...	x	y_0	x	x	x	y_2	...

x Missing label ○ Labels □ Instance data

Gomes, Grzenda, Mello, Read, Le Nguyen, Bifet. "A Survey on Semi-Supervised Learning for Delayed Partially Labelled Data Streams" *CSUR* (2022).

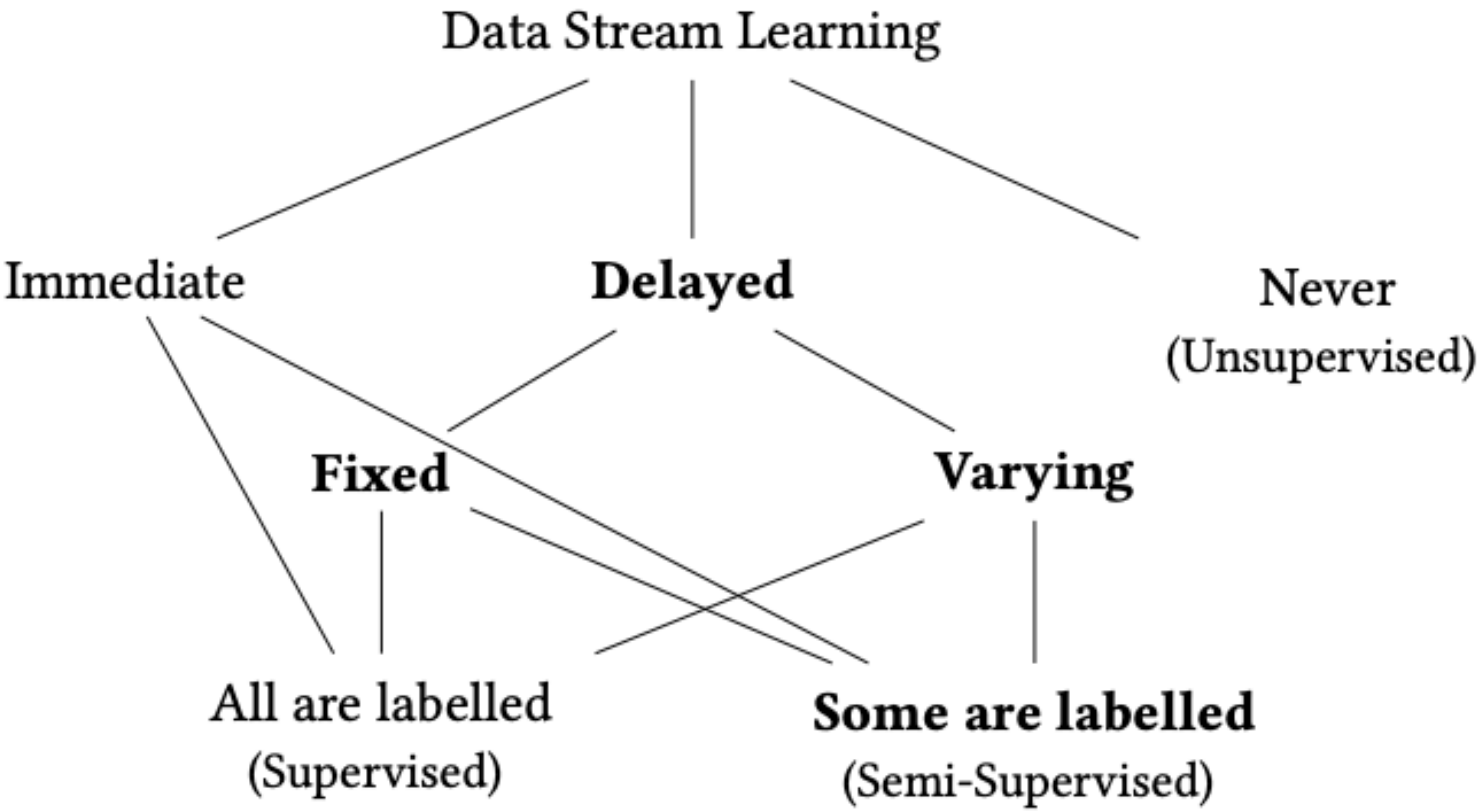
Semi-supervised learning

Supervised Learning + Unsupervised Learning

“Typically, ***semi-supervised learning*** *algorithms attempt to* ***improve performance in one of these two tasks by utilizing information generally associated with the other.*** *For instance, when tackling a classification problem, additional data points for which the label is unknown might be used to aid in the classification process. For clustering methods, on the other hand, the learning procedure might benefit from the knowledge that certain data points belong to the same class.” [Engelen and Hoos, 2020]*

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

Learning according to labels arrival time



Gomes, Grzenda, Mello, Read, Le Nguyen, Bifet. "A Survey on Semi-Supervised Learning for Delayed Partially Labelled Data Streams" *CSUR* (2022).

Related Problems (SSL in DS)

- Active learning
- Transductive learning
- Weakly multi-labelled data
- Initially Labeled Streaming Environment
- Few-shot learning
- Concept evolution
- Online Continual learning

* See section 2.1 from reference

Gomes, Grzenda, Mello, Read, Le Nguyen, Bifet. "A Survey on Semi-Supervised Learning for Delayed Partially Labelled Data Streams" *CSUR* (2022).

SSL adapted to Streams

- A practical approach to **adapt batch** techniques to streaming is through **sliding or tumbling windows**

- A general strategy for SSL is to use unlabelled examples to **build a representation of the input data**, and then use this representation as input to a model for obtaining predictions.

- The idea is that **an improved representation will lead to improved predictions**; and since representation learning can naturally be an unsupervised task, training labels are not required

- Examples: Restricted Boltzmann machines (RBMs), Auto-encoders, Cluster representations

Gomes, Grzenda, Mello, Read, Le Nguyen, Bifet. "A Survey on Semi-Supervised Learning for Delayed Partially Labelled Data Streams" *CSUR* (2022).

Evaluation of partially labeled data

- **Labeling ratio**, 1%, 5%, …, if too much data is labeled, then it is likely that a supervised algorithm would prevail
- **Delay simulation**, often simulated based on number of instances
- **Comparison** against **SSL and supervised** strategies

Avital Oliver, Augustus Odena, Colin A Raffel, Ekin Dogus Cubuk, and Ian Goodfellow. 2018. **Realistic Evaluation of Deep Semi-Supervised Learning Algorithms**. In Advances in Neural Information Processing Systems. 3238–3249.

Clustering Data Streams

Clustering data streams refers to grouping instances into clusters as the data continuously flows in, which normally includes **two phases**:

1. Online Step

a. Micro-Cluster* Formation: Incoming data points are incrementally processed and assigned to micro-clusters.

b. Micro-Cluster Maintenance: The micro-clusters are periodically updated as new data arrives. This includes adjusting the micro-cluster centroids and merging or splitting clusters based on defined thresholds.

2. Offline Step

Periodically or upon request, micro-clusters are aggregated into macro-clusters (or simply clusters) to provide a higher-level view of the data.

* **Micro-clusters** are small, temporary “clusters” that capture local density information and are typically represented by statistical summaries like centroid, weight, and radius.

Bifet, Albert, Ricard Gavalda, Geoffrey Holmes, and Bernhard Pfahringer.
Machine learning for data streams: with practical examples in MOA. MIT press, 2023.

CluStream

1. Online Step

For each new **instance *i*** that arrives:

- ***i*** is **absorbed** by a micro-cluster
- ***i* starts** a new **micro-cluster** of its own
  - **Delete oldest** micro-cluster
  - **Merge** two of the **oldest** micro-cluster

2. Offline Step

Apply k-means using micro-clusters as instances

Aggarwal, Charu C., S. Yu Philip, Jiawei Han, and Jianyong Wang. "A framework for clustering evolving data streams." In *Proceedings 2003 VLDB conference*, pp. 81-92. Morgan Kaufmann, 2003.

Anomaly Detection (streams)

- Anomaly detection for data streams requires the identification of abnormal data patterns in a continuous flow of data

- **Challenges**
  - Adapting to concept drifts without missing out on anomalies
  - Detecting rare anomalies amidst high-volume data streams
  - And more…

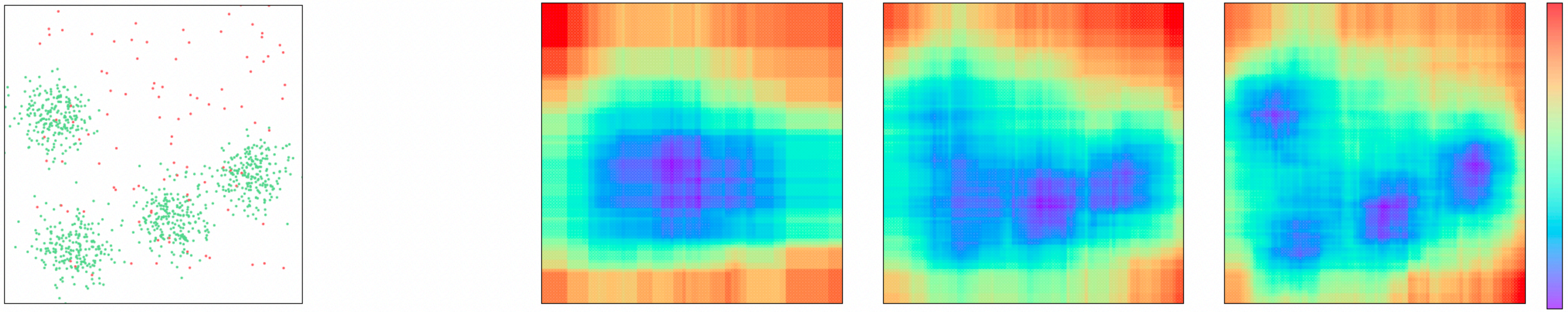
Online Isolation Forest

- Inspired by the classic Isolation Forest
- models the data distribution via an ensemble of multi-resolution histograms endowed with a dynamic mechanism to learn new points and forget old ones
- Each histogram collects the points count within its bins and, upon reaching a maximum height, a bin undergoes a split to increase the resolution of the histogram in the most populated regions of the space. Conversely, bins in sparsely populated regions are eventually aggregated, thereby decreasing the

Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. "Isolation forest." ICDM, 2008.

Filippo Leveni, G W Cassales, B Pfahringer, A Bifet, and G Boracchi. "Online Isolation Forest." ICML, 2024

Online Isolation Forest



(a) Data stream $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t \in \mathbb{R}^d$.

(b) Anomaly scores s at different time instants t, from left to right.

- Genuine data (green) are more densely distributed than anomalous data (red)

- OIF processes each data point individually online, assigning an anomaly score to each

- As more data is available, OIF continuously updates and refines the anomaly scores based on the evolving data distribution.

Filippo Leveni, G W Cassales, B Pfahringer, A Bifet, and G Boracchi. "Online Isolation Forest." ICML, 2024

Practical example

IJCAI_2024_advanced.ipynb

Coming up next in 2024

- **Upcoming Tutorials**
  - PAKDD: May 2024 (Taipei, Taiwan) **[done!]**
  - IJCAI: August 2024 (Jeju, South Korea) **[this one]**
  - KDD: August 2024 (Barcelona, Spain)
  - Kiwi Pycon: August 2024 (Wellington, New Zealand)
  - ECML: September 2024 (Vilnius, Lithuania)
- **CapyMOA next release** (August 2024)

Conclusion

- Streaming data is everywhere — often in a delayed and partially labeled
- ML algorithms for data streams should be **accurate**, **adaptive** and **efficient**
- **CapyMOA** can be easily extended for many stream tasks

Contact: heitor.gomes@vuw.ac.nz

Discord

https://discord.gg/RekJArWKNZ

Thank you!

https://github.com/adaptive-machine-learning/CapyMOA