Introduction to Learning from Streaming Data

KEEPER Workshop Tutorial 2024

Nuwan Gunasekara*[1], Sepideh Pashami [1],
https://nuwangunasekara.github.io/KEEPER2025/

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

https://capymoa.org/

[1] Center for Applied Intelligent Systems Research, Halmstad University, Sweden

Anomaly Detection

Anomaly Detection for Data Streams

- Identification of anomalous data 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 (OIF)

- Inspired by the classic Isolation Forest [1]

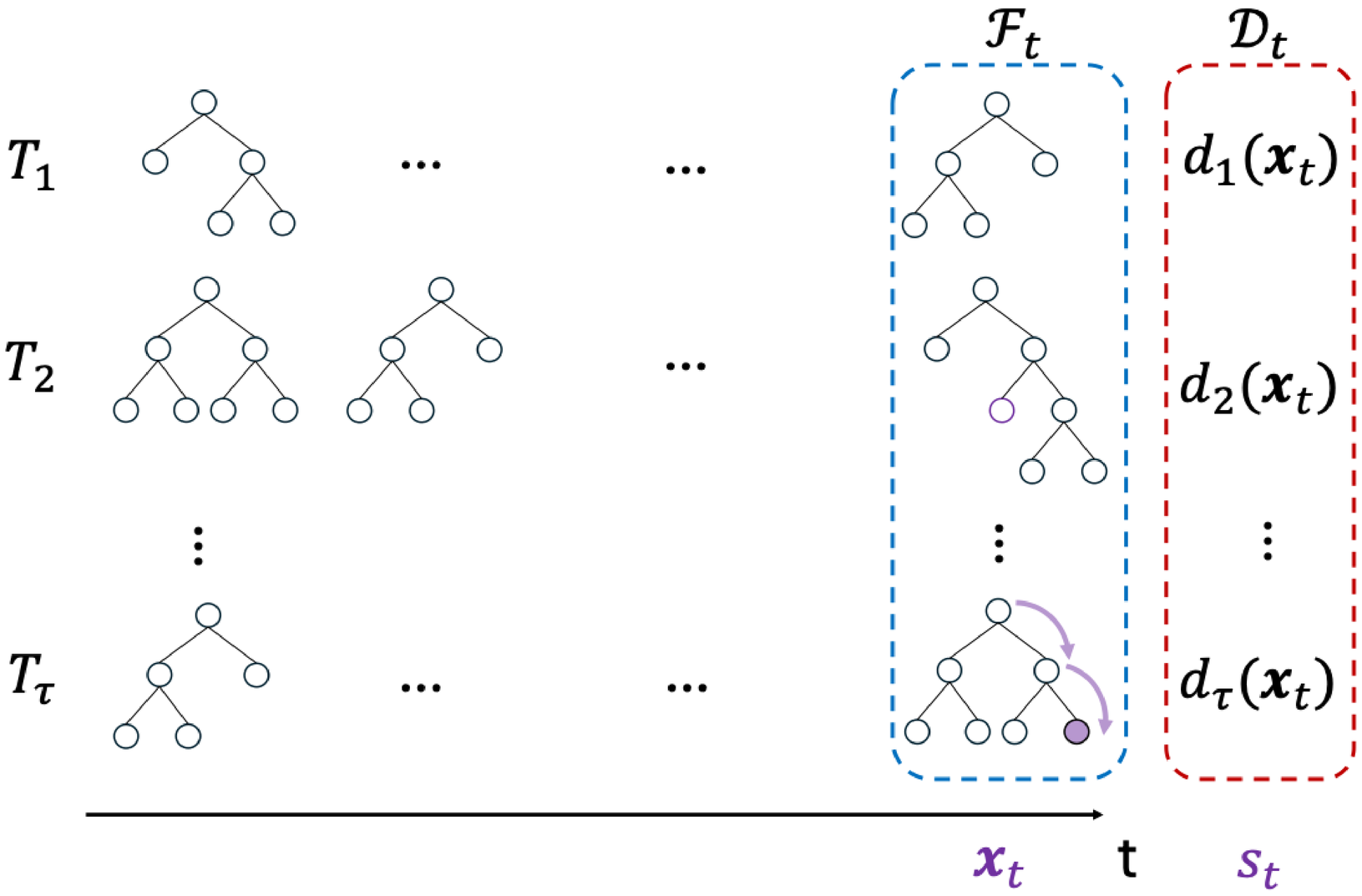
- OIF uses a **group of histograms** at **different levels of detail** to **capture** the data patterns, with a flexible system that can **learn from new data and gradually forget older data**.

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

[2] Filippo Leveni, G W Cassales, B Pfahringer, A Bifet, and G Boracchi. "Online Isolation Forest." ICML, 2024
https://icml.cc/virtual/2024/poster/34674

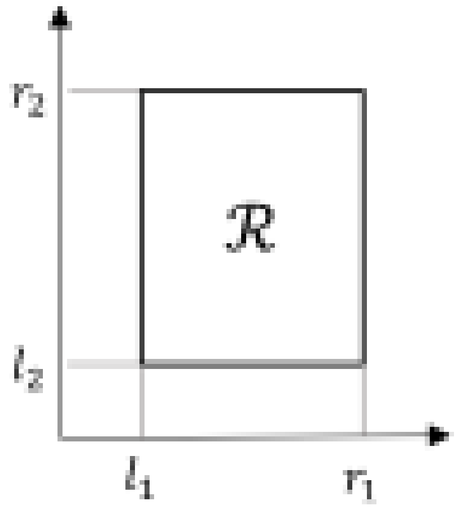
Online Isolation Forest

Anomaly Scoring

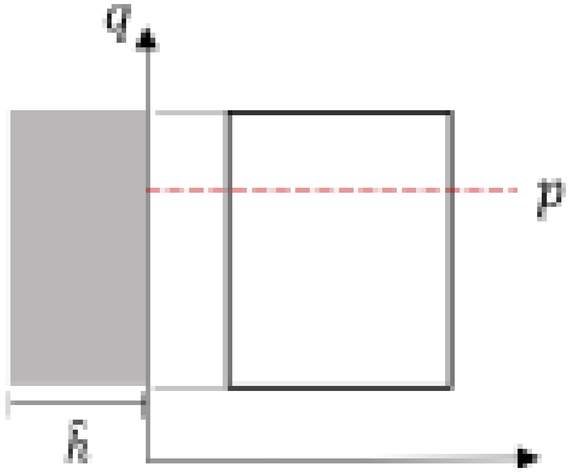


Online Isolation Forest

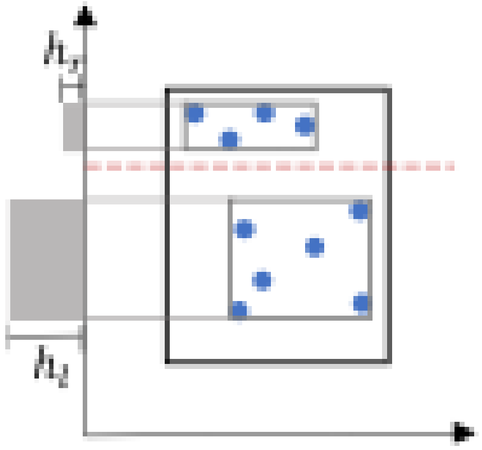
Splitting



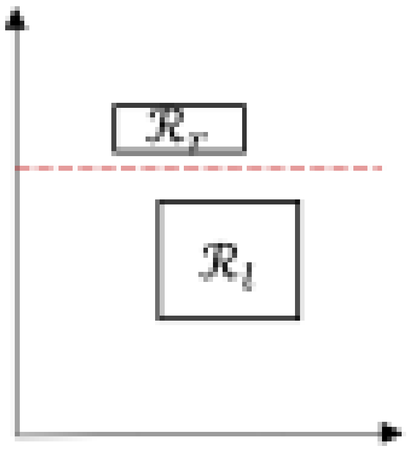
(a) Bin support $\mathcal{R}$ and its boundaries $[l_i, r_i]$.



(b) Maximum bin height $\widehat{h}$ and split information q and p.



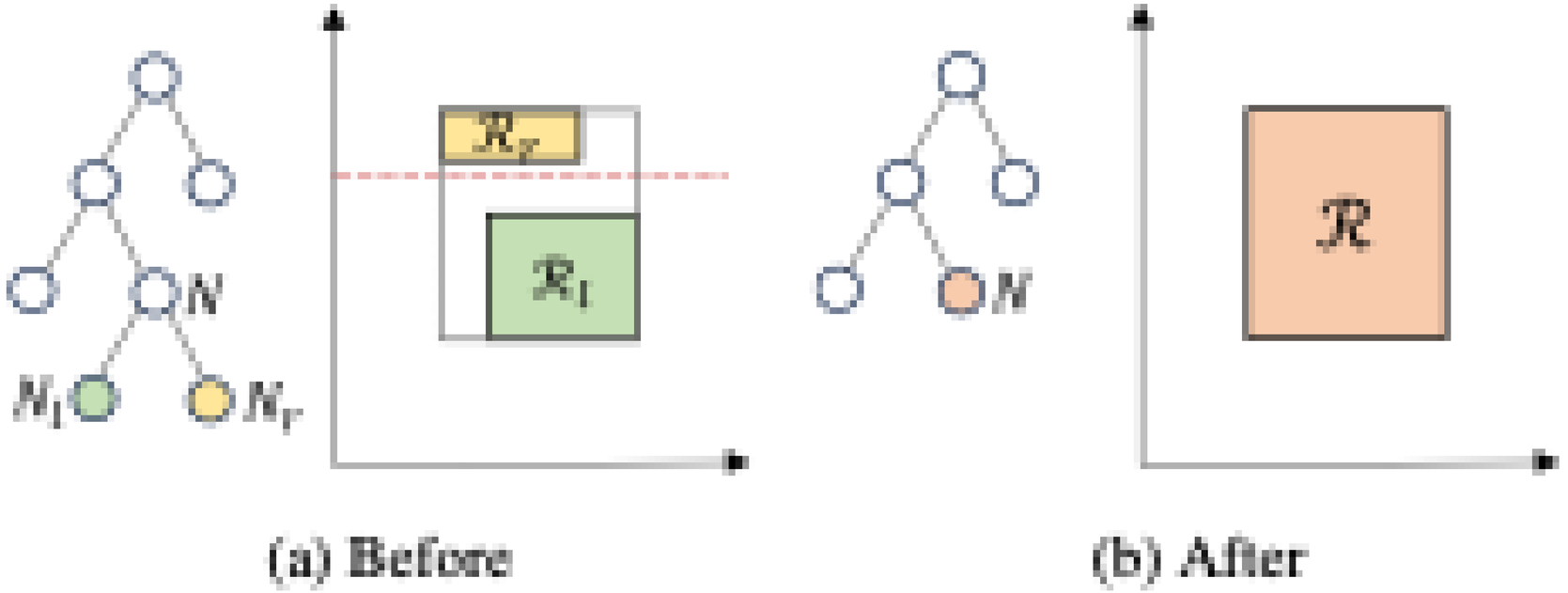
(c) Sampled points $\mathcal{X}$ and new bins height h_l and h_r.



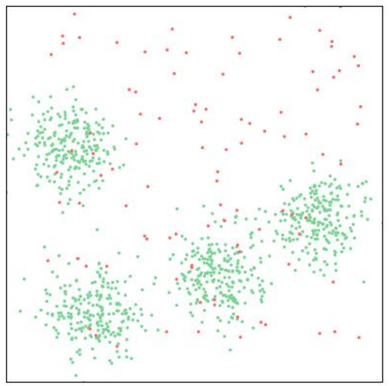
(d) New bins support $\mathcal{R}_l$ and $\mathcal{R}_r$.

Online Isolation Forest

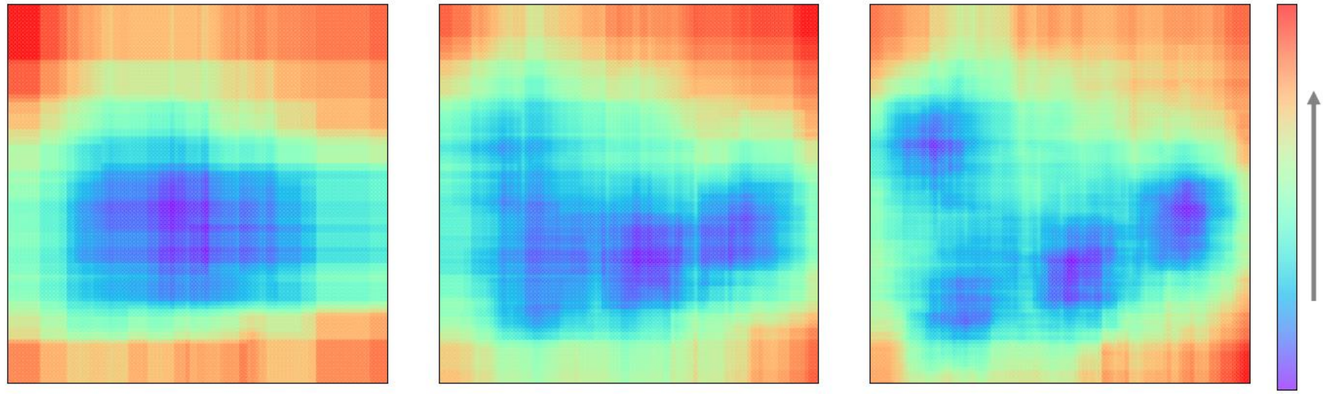
Forgetting



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

03_KEEPER2025_anomaly_detection.ipynb

Coming up next in 2025

- **Upcoming Tutorials**
  - PAKDD: May 2024 (Taipei, Taiwan) **[done!]**
  - IJCAI: August 2024 (Jeju, South Korea) **[done!]**
  - KDD: August 2024 (Barcelona, Spain) **[done!]**
  - Kiwi Pycon: August 2024 (Wellington, NZ) **[done!]**
  - ECML: September 2024 (Vilnius, Lithuania) **[done!]**
  - ICDE: May 2025 (Hong Kong, China)
  - PAKDD: June 2025 (Sydney, Australia)

- **CapyMOA next release** (Early May 2025)

Conclusion

- Streaming data is everywhere
- 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