UNIVERSITÀ DEGLI STUDI DI SALERNO
DIPARTIMENTO DI INFORMATICA

IoT Data Analytics

Examples of Projects

L. Caruccio, AY 2024-25

Batch vs Online Learning

- Comparison of **batch learning vs online learning** models

- The goal is to evaluate how these models perform in:
  - **stationary scenarios**
  - **scenarios with concept drift**, where the underlying data distribution changes over time.

- Several online learning algorithms are provided by **different libraries** (River, scikit-multiflow, Vowpal Wabbit)

Batch vs Online Clustering

- Comparison between **static and online clustering methods and online clustering algorithms** when applied to streaming data scenarios.

- Comparison in terms of **clustering quality** and **computational efficiency**
  - Evaluation of the trade-offs involved in using batch vs online clustering strategies

- The **River** library provides different online algorithms

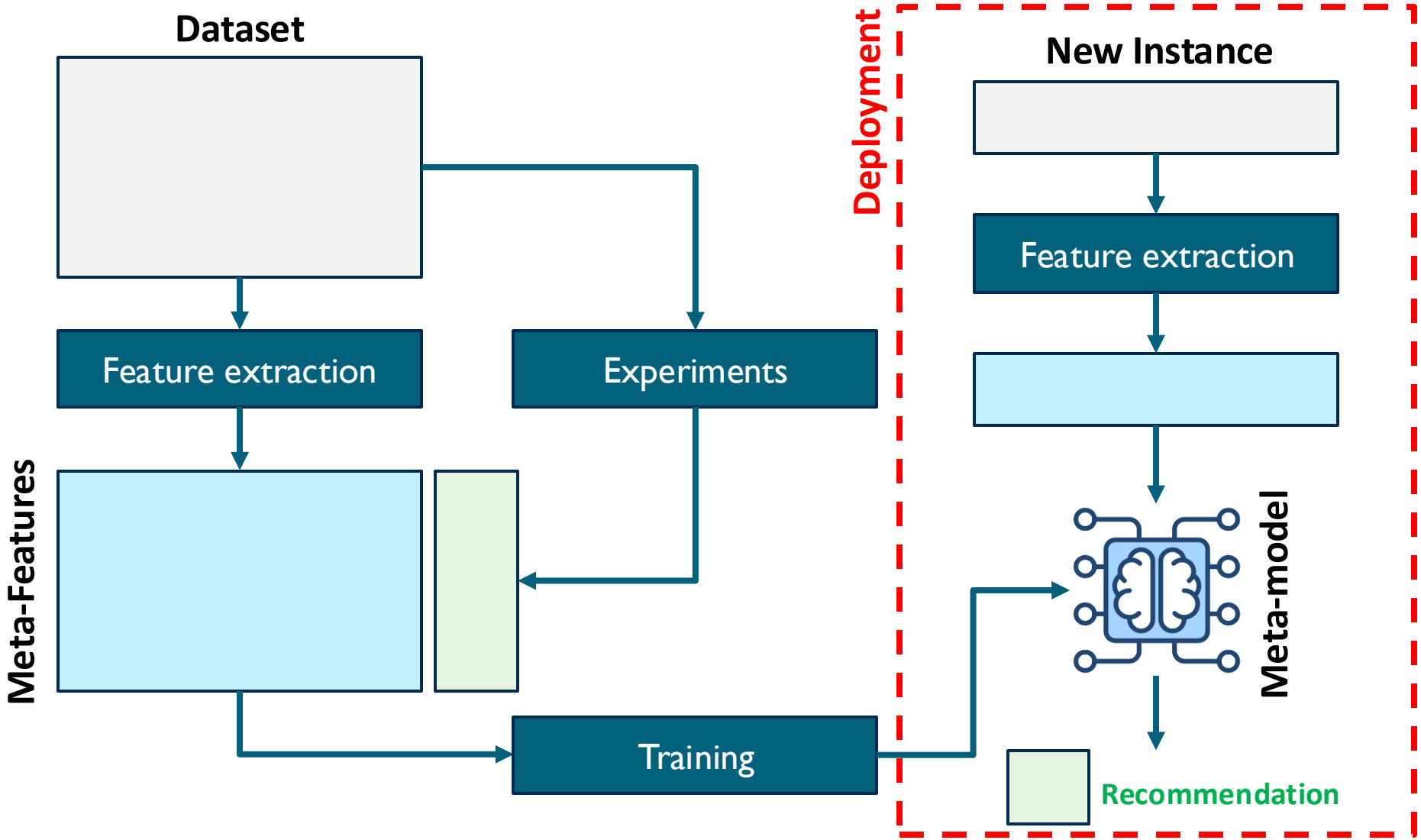
Meta Learning

- **Meta-learning** refers to training a machine learning model on a dataset composed of **metadata**

- The trained meta-model **can provide recommendations basing on the learned past experience**
  - (e.g., recommending the best parameters, models etc.)

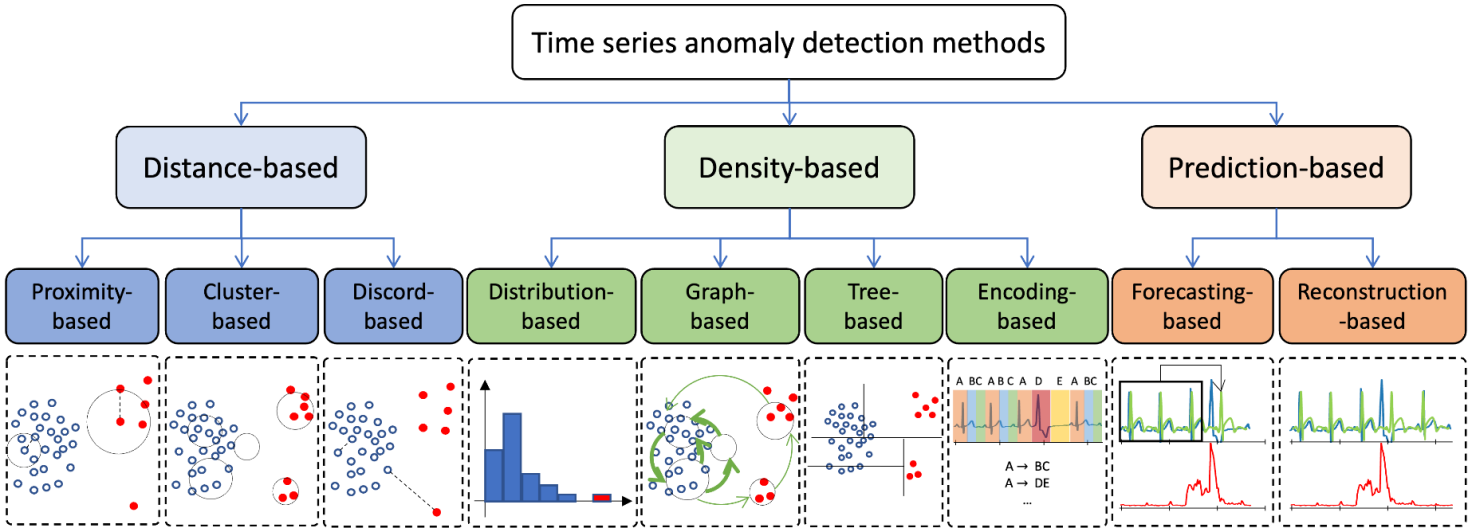
- This approach has been recently used to provide recommendations **based on the input characteristics**

Meta Learning: how it works



Meta Learning: Anomaly Detection

- Implementation of a meta-learning approach to identify the best anomaly detector for a given series



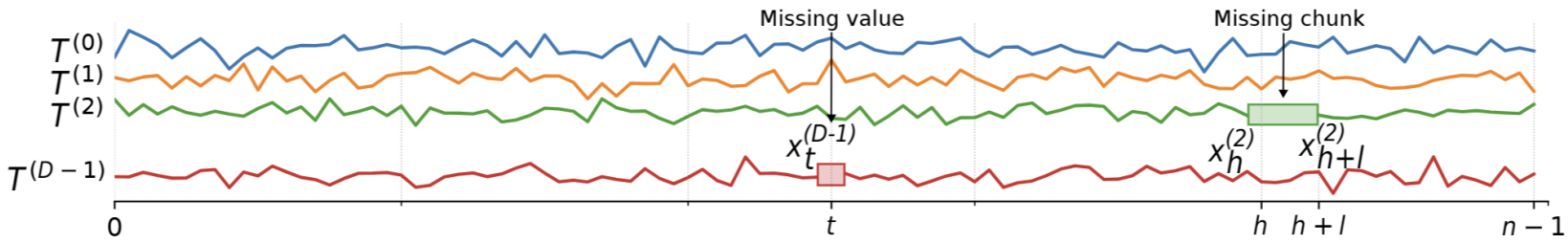
- The meta-model recommends the best model to use basing on the time series features

Meta Learning: Anomaly Detection

- Comparison of different feature extraction techniques
  - tsfresh
  - tsfel
  - tsfeatures
  - …

- Evaluating ⏳ Efficiency vs 🎯 Accuracy
  - Use of all features
  - Selecting the best features
  - Aggregating groups of features

Meta Learning: data imputation



- Missing values are common in time series data, as they contain data usually collected from sensors
  - The presence of missing values significantly degrades the performance of any downstream task

- Data imputation refers to filling missing values by identifying potential candidates within the series

Meta Learning: data imputation

- There are several approaches for imputing time series data
  - Interpolation
  - Propagation
  - Deep-Learning
  - …
- The performance of data imputers varies significantly according to the series' characteristics
- Development of a meta-model to recommend **the best imputer for a given series**

Resources

- Several **time series repositories** are available online.

- One of the most comprehensive ones is **TSB-UAD** (https://github.com/TheDatumOrg/TSB-UAD) which is focused on **univariate time series:**
  - It contains **12686 time series with labeled anomalies** spanning different domains with high variability of anomaly types, ratios, and sizes.

- **TSB-AD-M**, instead, is a recent repository containing multivariate time series data (Download link: https://www.thedatum.org/datasets/TSB-AD-M.zip)

Resources

- Online learning/clustering libraries:
  - River: https://riverml.xyz/latest/
  - Scikit-Multiflow: https://scikit-multiflow.github.io/
  - Vowpal Library: https://vowpalwabbit.org/index.html
- Time series imputation libraries:
  - PyPots: https://github.com/WenjieDu/PyPOTS (Tutorials available at https://github.com/WenjieDu/BrewPOTS)
- Time series Feature Extraction libraries:
  - tsfresh: https://tsfresh.readthedocs.io/en/latest/
  - tsfel: https://tsfel.readthedocs.io/en/latest/
  - tsfeatures: https://github.com/Nixtla/tsfeatures
- Clustering evaluation metrics:
  - https://www.geeksforgeeks.org/clustering-metrics/

Resources

- Articles on Google Scholar:
  - Examples of meta-learning approaches
    - Tavares, G. M., & Barbon, S. J. (2023). Matching business process behavior with encoding techniques via meta-learning: An anomaly detection study. *Computer Science and Information Systems*, *20*(3), 1207-1233.
    - Sylligardos, E., Boniol, P., Paparrizos, J., Trahanias, P., & Palpanas, T. (2023). Choose wisely: An extensive evaluation of model selection for anomaly detection in time series. Proceedings of the VLDB Endowment, 16(11), 3418-3432.
  - Time series imputation surveys
    - Lepot, M., Aubin, J. B., & Clemens, F. H. (2017). Interpolation in time series: An introductive overview of existing methods, their performance criteria and uncertainty assessment. *Water*, *9*(10), 796.
    - Wang, J., Du, W., Yang, Y., Qian, L., Cao, W., Zhang, K., ... & Wen, Q. (2024). Deep learning for multivariate time series imputation: A survey. *arXiv preprint arXiv:2402.04059*.
  - Online learning approaches survey
    - Hoi, S. C., Sahoo, D., Lu, J., & Zhao, P. (2021). Online learning: A comprehensive survey. *Neurocomputing*, *459*, 249-289.
  - Time series anomaly detection approaches survey
    - Boniol, P., Liu, Q., Huang, M., Palpanas, T., & Paparrizos, J. (2024). Dive into time-series anomaly detection: A decade review. arXiv preprint arXiv:2412.20512.