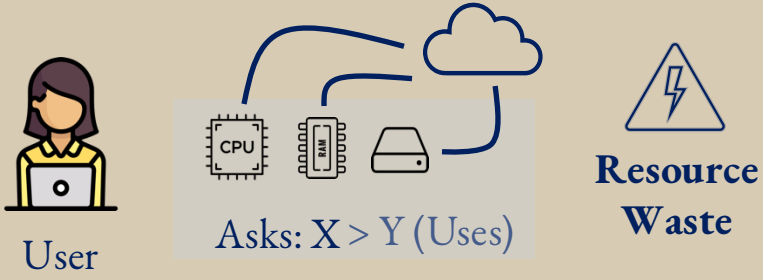
Smart and Sustainable Cloud Resource Management with Minimal Use of Machine Intelligence

Georgia Christofidi
Supervisor: Thaleia Dimitra Doudali
@**EuroDW'25**, March 31st

Current Problems in Cloud Resource Management

Challenge 1: Low Resource Efficiency in the Cloud.



Challenge 2: Increased Carbon Emissions due to **exponential growth** of Computing.

Key drivers:

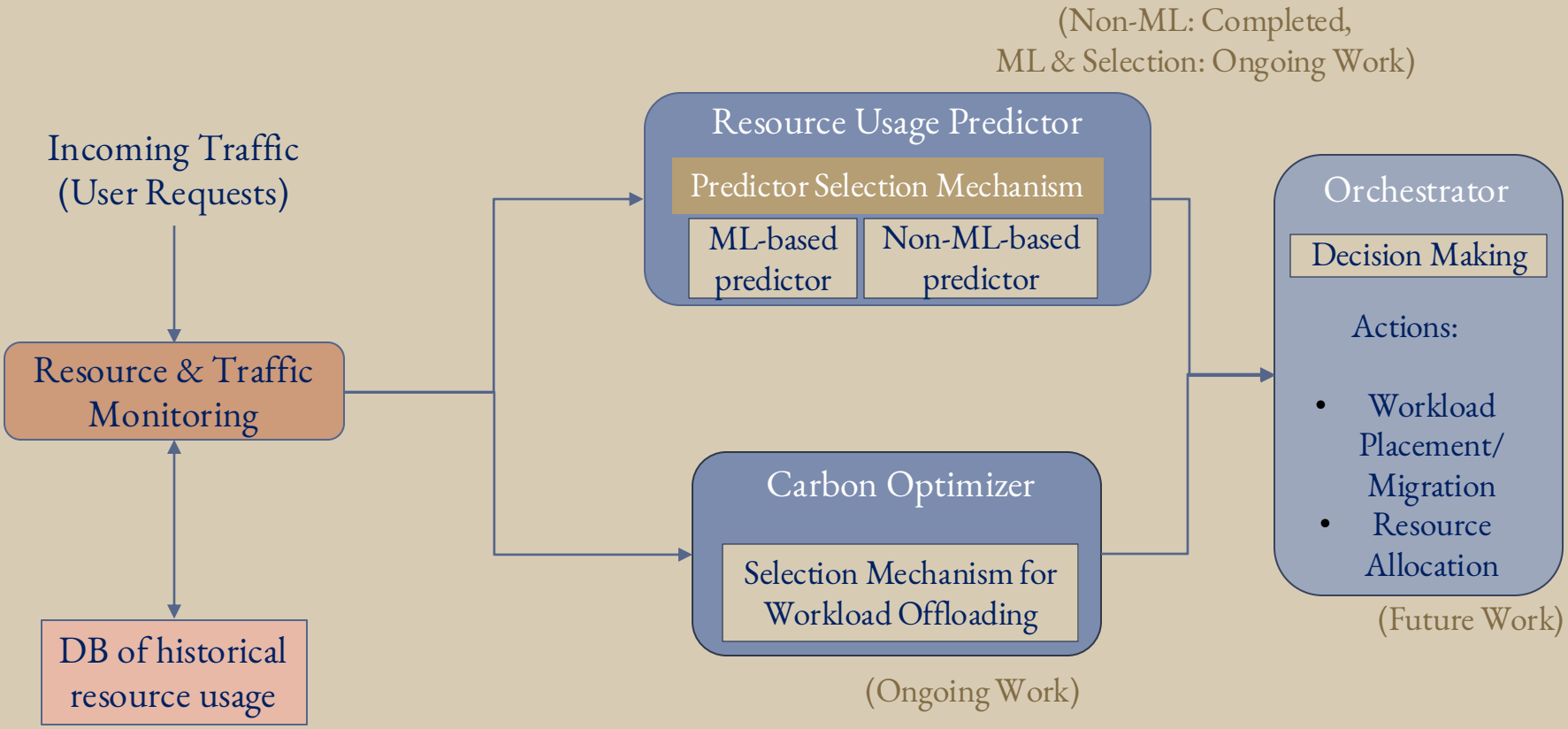
- ML applications
- Generative AI
- Video streaming

AI Model	Carbon Impact of Training*	Real-word equivalent example
GPT-3	500 metric tons of CO2eq.[1]	500 round-trip flights from Madrid to New York for one passenger.
GPT-4	12,456 - 14,994 metric tons CO2eq (*estimated*).[2]	50-60 fully loaded Boeing 747 flights.

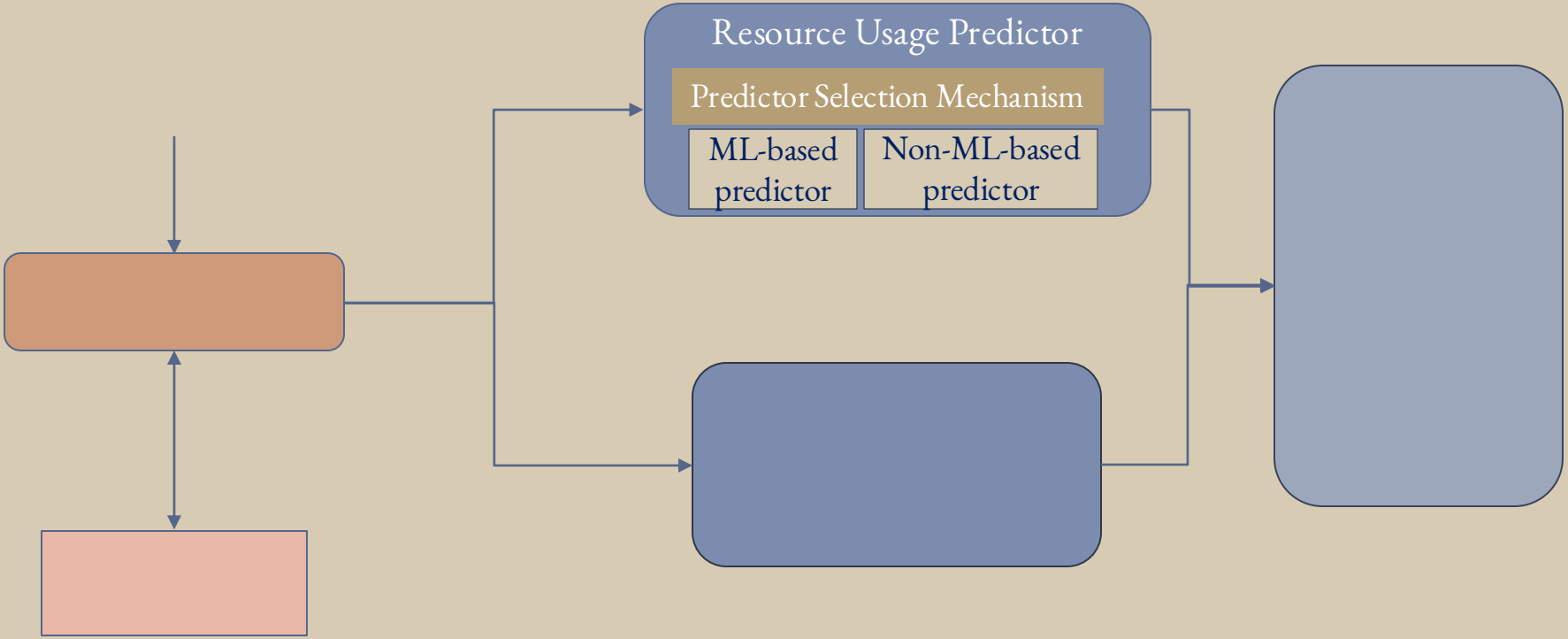
*Training only accounts for 43% of lifecycle carbon emissions. [1]

Sources [1]: Beyond Efficiency: Scaling AI Sustainably
[2]: https://towardsdatascience.com/the-carbon-footprint-of-gpt-4-d6c676eb21ae

Thesis Overview – System Design



Resource Usage Predictor Component



Resource Usage Predictor Component

Approach: Characterize Datasets of Resource Usage in the Cloud.*

Exploration of **public open-source** datasets across different:

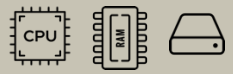
Cloud providers
Resource Types
Resource Levels

Takeaway: Resource usage is highly correlated in time.

In such cases, **ML**-based forecasting may **not** be **necessary** and **simple predictive models** can deliver the desired levels of accuracy.

Data correlation in time depends on:

Resource Types



Resource Levels

Workload

Physical-Virtual Machine

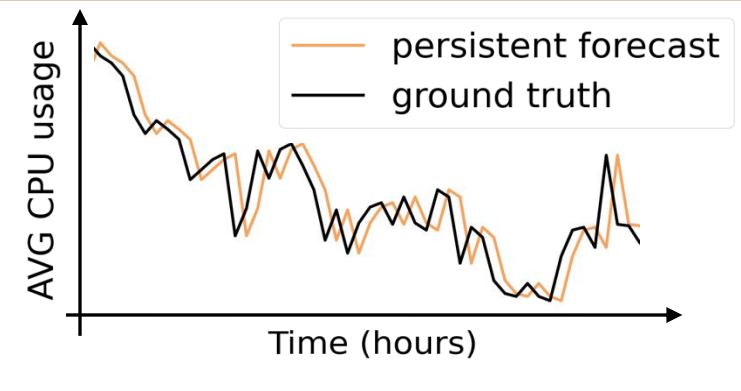
Measurement

MIN, MAX, AVG

*Georgia Christofidi, Konstantinos Papaioannou, and Thaleia Dimitra Doudali. 2023. Is Machine Learning Necessary for Cloud Resource Usage Forecasting? In Proceedings of the 2023 ACM Symposium on Cloud Computing (SoCC '23). Association for Computing Machinery.

Predictor Component– non-ML-based (Completed)

Proposed Solution Across Usecases: Persistent Forecast.*



Predicted Value(t) = Ground Truth(t – 5 mins)

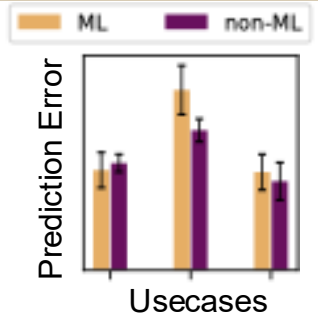
Takeaway: Persistent Forecast is **highly accurate** for future resource usage in the cloud at the **physical and the virtual machine level.** (AVG Prediction Error < 6%).

*Georgia Christofidi, Konstantinos Papaioannou, and Thaleia Dimitra Doudali. 2023. Is Machine Learning Necessary for Cloud Resource Usage Forecasting? In Proceedings of the 2023 ACM Symposium on Cloud Computing (SoCC '23). Association for Computing Machinery.

Predictor Component– ML-based (Ongoing)

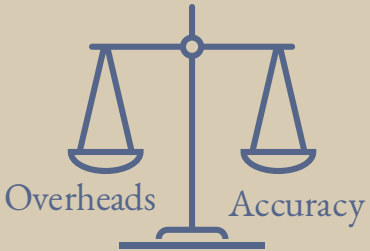
ML Model: Long Short-Term Memory Neural Networks (LSTMs)*.

~ Accuracy
↑↑↑ Overheads



If LSTMs have similar accuracy to the non-ML methods, do we really **need** them?

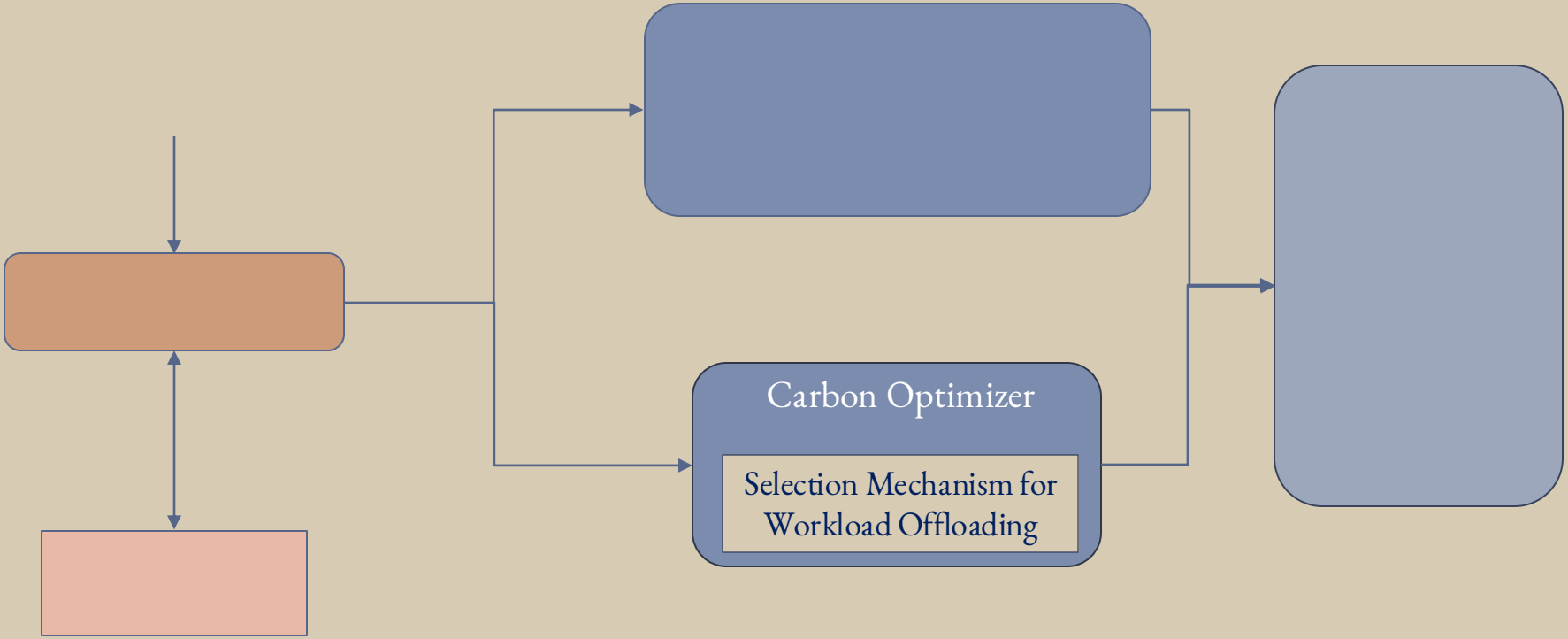
Future Work: Explore more ML-based Methods.



Takeaway: To complement the non-ML predictor, we need a **highly accurate** predictor with **low overheads**.

*Georgia Christofidi, Konstantinos Papaioannou, and Thaleia Dimitra Doudali. 2023. Is Machine Learning Necessary for Cloud Resource Usage Forecasting? In Proceedings of the 2023 ACM Symposium on Cloud Computing (SoCC '23).
Toward Pattern-based Model Selection for Cloud Resource Forecasting. In Proceedings of the 3rd Workshop on Machine Learning and Systems (EuroMLSys '23). Association for Computing Machinery.

Carbon Optimizer Component



Carbon optimizer component (Ongoing)

Approach: Characterize the carbon savings of migration of different applications.*

Usecase: Two microservice applications hosted by a national company.

App (Location)	Carbon (mgCO2eq)
ES Social Network	72.72
SE Social Network	7.06
ES Media Streaming	166.17
SE Media Streaming	16.13

↑ computationally demanding
↑ impact in sustainability

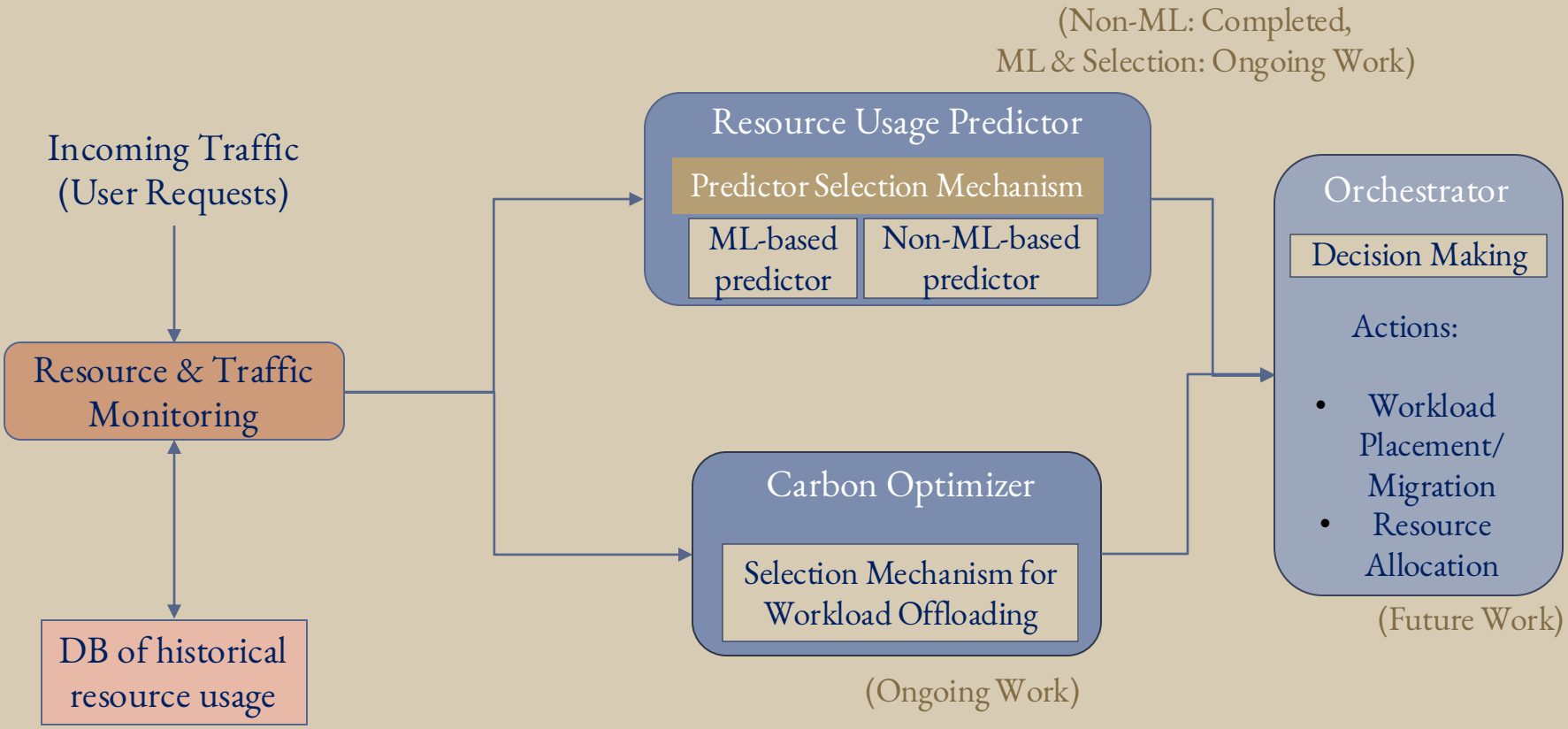
Idea: Offload workloads from Spain [E][S] to Sweden [S][E] (10x less carbon intensive).

Takeaway: We need a **smart, application-specific solution** for selecting workloads for offloading.

Future Work: Extend to more application and workload types.

* Christofidi, Georgia, et al. "CaRE: Towards Carbon and Resource Efficient Orchestration at the Cloud-Edge

Thesis Overview – System Design



Publications

- **CaRE: Towards Carbon and Resource Efficient Orchestration at the Cloud-Edge Continuum**

Georgia Christofidi, Francisco Alvarez Terribas, Jesus Alberto Omana Iglesias, Nicolas Kourtellis, Thaleia-Dimitra Doudali. Machine Learning for Edge-Cloud Systems (ML4ECS). In conjunction with **HiPEAC 2025**.

- **Augmenting Cloud Resource Management with the Necessary Amount of Machine Intelligence.**

Georgia Christofidi, Thaleia-Dimitra Doudali. 30th International European Conference on Parallel and Distributed Computing (**EuroPar ′24**, PhD Symposium).

- **Do Predictors for Resource Overcommitment Even Predict?**

Georgia Christofidi, Thaleia Dimitra Doudali. In Proceedings of the 4th Workshop on Machine Learning and Systems (EuroMLSys ’24). In conjunction with the 2024 European Conference on Computer Systems (**EuroSys ’24**)

- **Is Machine Learning Necessary for Cloud Resource Usage Forecasting?**

Georgia Christofidi, Konstantinos Papaioannou, Thaleia Dimitra Doudali. In Proceedings of the 14th Symposium of Cloud Computing (**SoCC ’23**)

- **Toward Pattern-based Model Selection for Cloud Resource Forecasting**

Georgia Christofidi, Konstantinos Papaioannou, Thaleia Dimitra Doudali. In Proceedings of the 3rd Workshop on Machine Learning and Systems (EuroMLSys ’23). In conjunction with the 2023 European Conference on Computer Systems (**EuroSys ’23**).