Is Machine Learning *Necessary* to Use in Cloud Resource Management?

Thaleia Dimitra Doudali

Assistant Professor

IMDEA Software Institute, Madrid, Spain

Monday, 22 April 2024

Talk at the ESwML workshop of EuroSys 2024

About Me

2015

2021

Start: October 2021

Born and raised in Greece.

Undergrad in ECE at NTUA, Athens, Greece.

PhD in CS at Georgia Tech, Atlanta, USA.

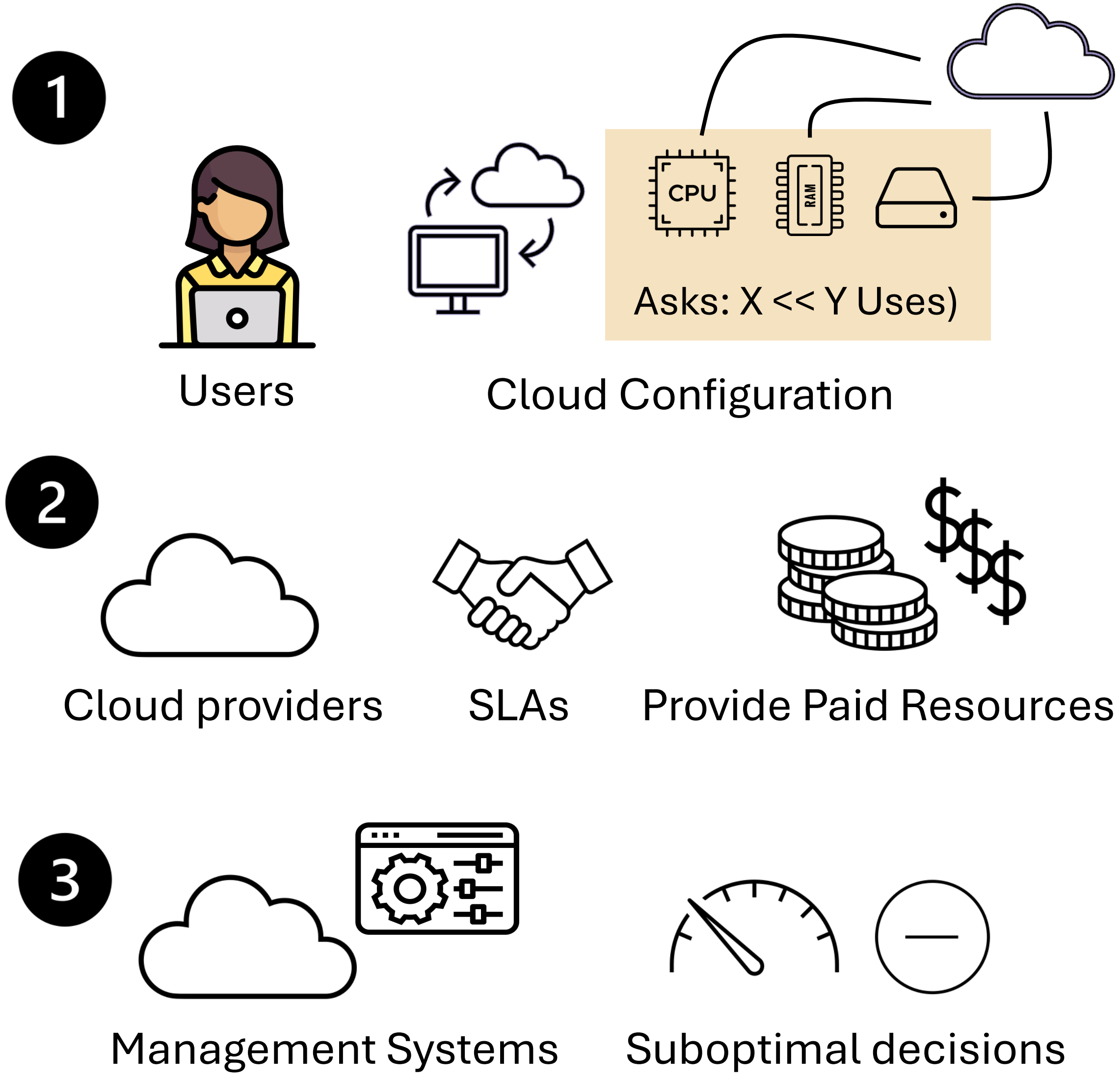
Advised by Ada Gavrilovska.

Assistant Professor at IMDEA, Madrid, Spain.

Website: https://thaleia-dimitradoudali.github.io/

Cloud Resource Efficiency

What causes low resource efficiency?

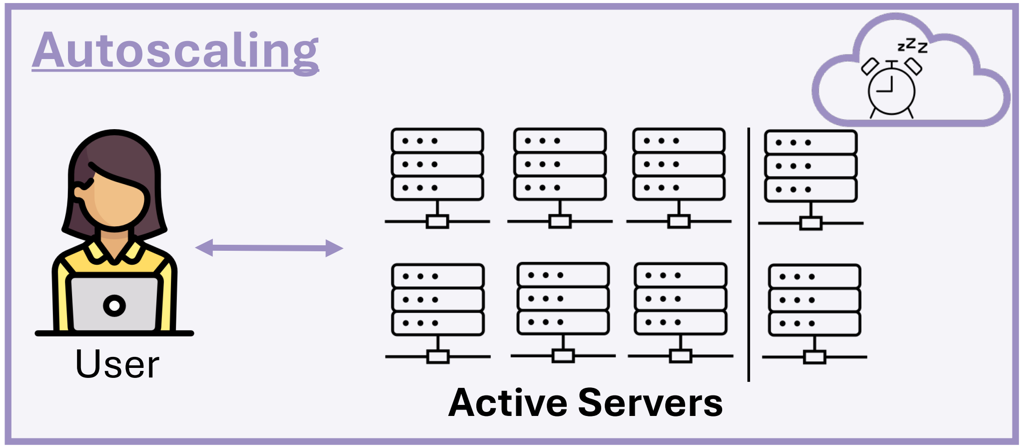


Data center resource utilization ~20%.

Cloud Resource Management Techniques

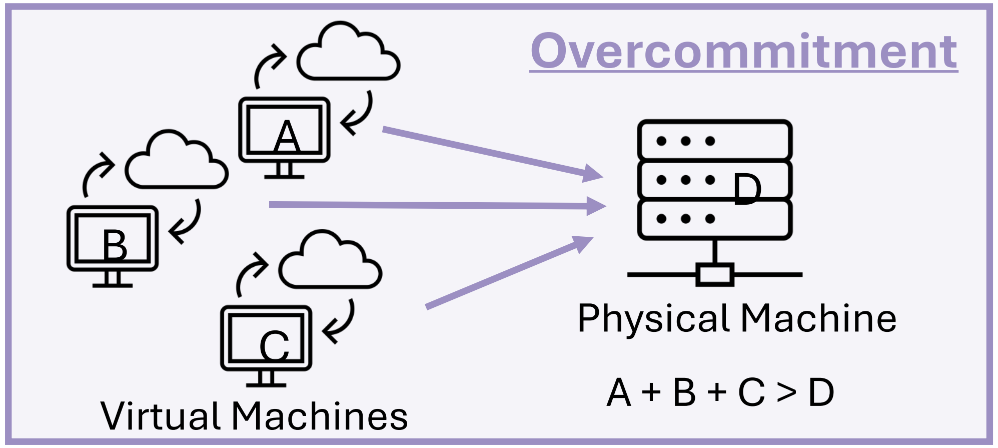
The following techniques help increase resource utilization and efficiency.

Basic idea: don't give to the user what they ask for, only what they actually use.



Dynamically **scale up or down** the number of computational resources e.g., active servers, number of CPUs.

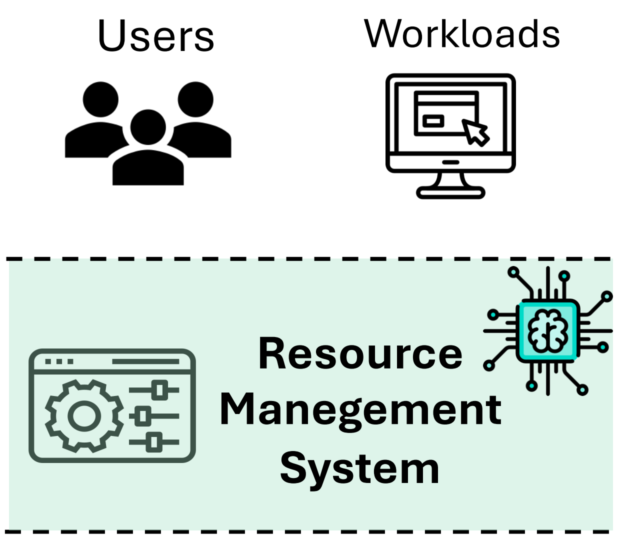
Delayed adjustment.
Performance reduction.



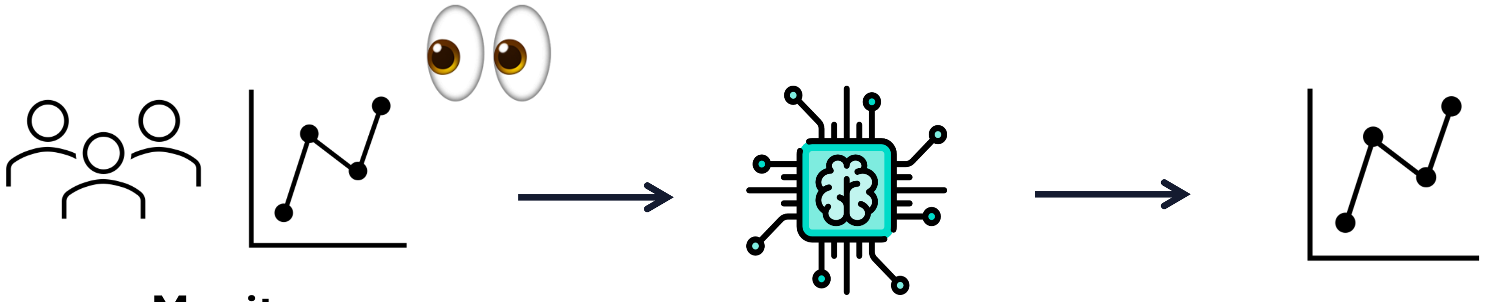
Allocate **more virtualized** resources than the ones physically available.
Assumes users *underutilize* resources.

Resource contention.
Performance reduction.

The Key: Resource Usage Forecasting



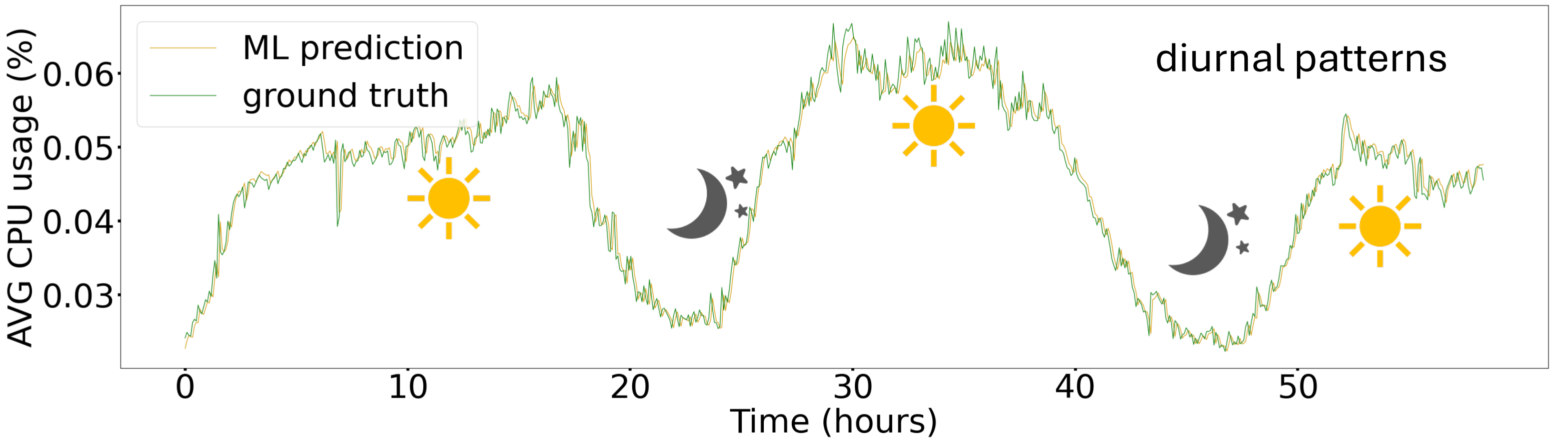
Hardware in the “Cloud”



Monitor
User Behavior
Workload Patterns

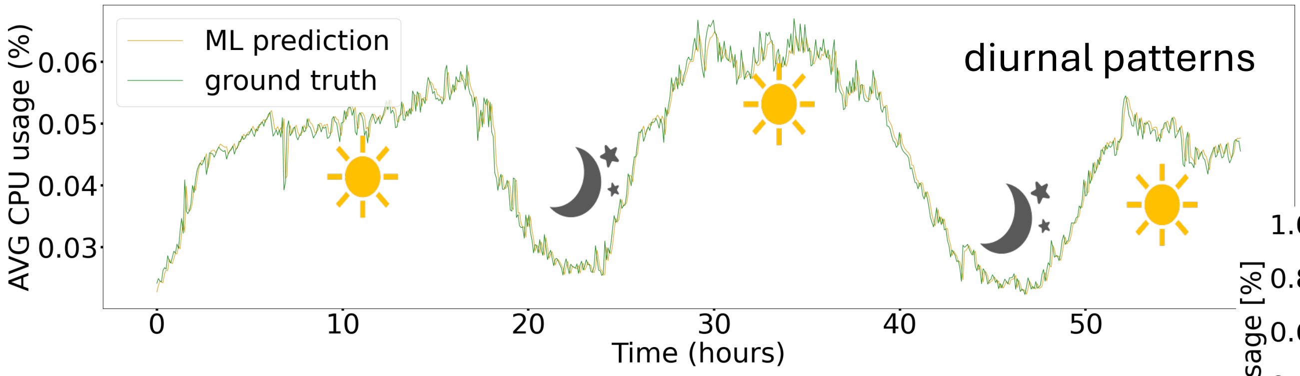
Predictor

Future
resource usage,
demand, load.

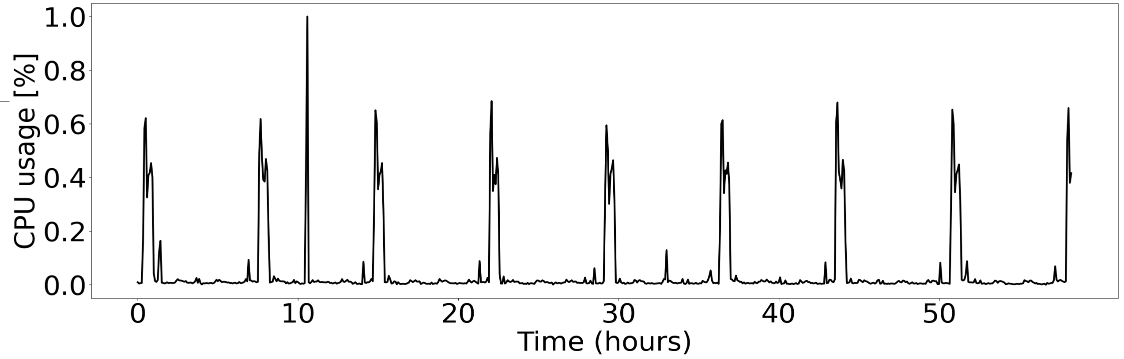


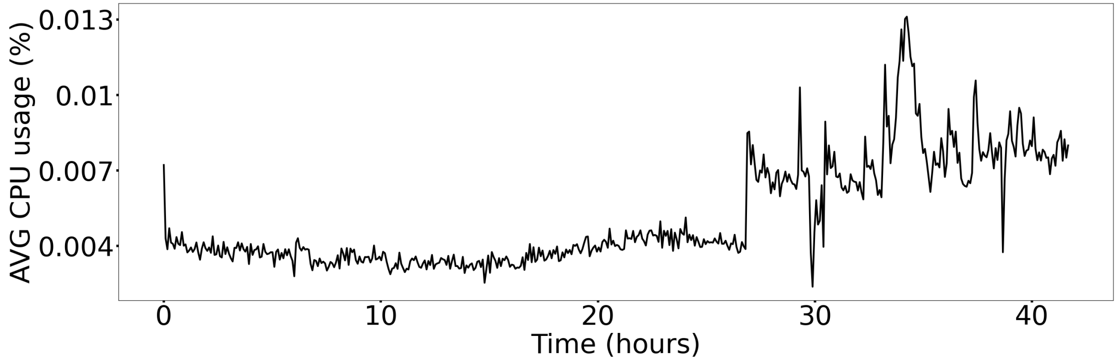
Accurate Predictiors → Timely and Effective Resource Management → Resource Efficiency.

Accurate Resource Usage Forecasting is Challenging

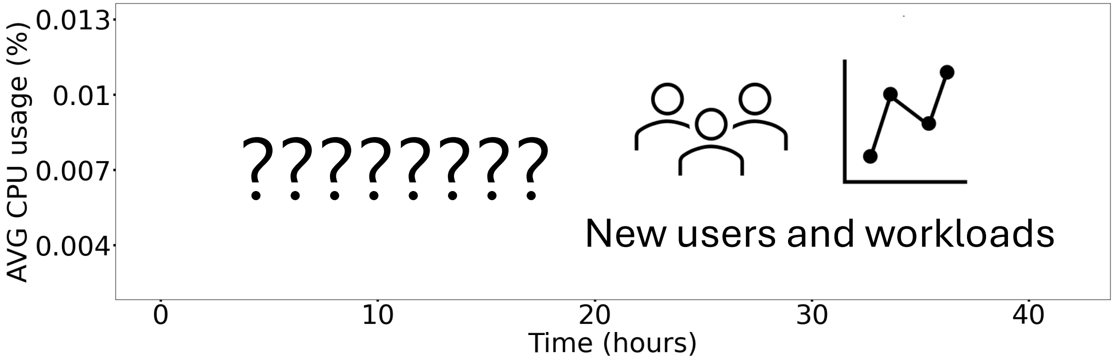


Stable, periodic, diurnal patterns are **predictable.**





Sudden changes, spikes, high dynamicity, are **hard to predict.**



Unseen patterns could be completely **unpredictable.**

Predictors for Resource Overcommitment

Existing Predictors
Future Usage =

1. Borg
90% * **Limit**

2. Resource Central
sum of the 99-th%-ile

3. N-Sigma
U+$N*std$(**U**)

4. Take-it-to-the-limit (TITTL) = Max (1, 2, 3)

Why Max?

To eliminate potential *under*-estimations, which may cause:

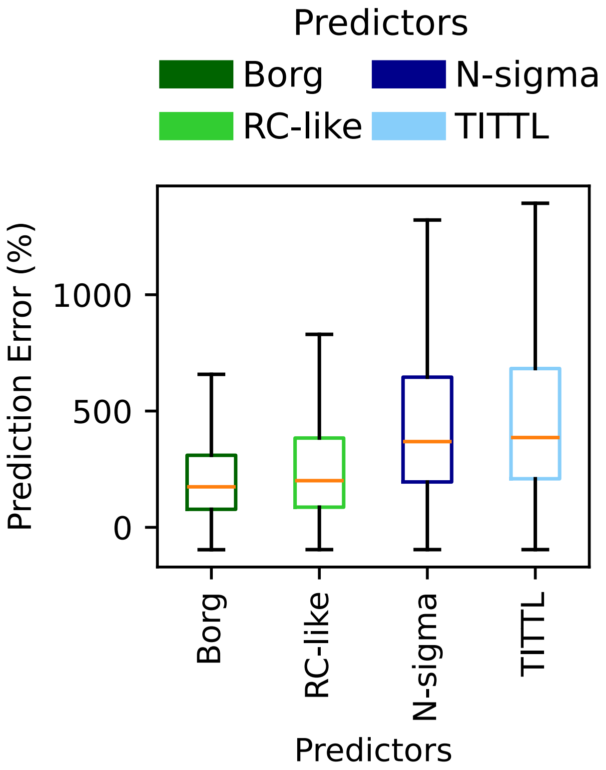
- Degraded workload performance.
- Unecessary resource auto-scaling.
- User SLA violations.

Simple, lightweight, explainable and easy to engineer in production-level.

Do they accurately predict resource usage or just protect from under-estimations??

[EuroMLSys '24] *Do Predictors for Resource Overcommitment Even Predict?* Georgia Christofidi and Thaleia Dimitra Doudali.

Do they Even Predict?



😱 Prediction error is extremely high, especially for TITTL.

😱 **Predicted resource usage >> resource limit.**
🙅 The cloud provider will NEVER allocate to a user more resources than requested and paid for!

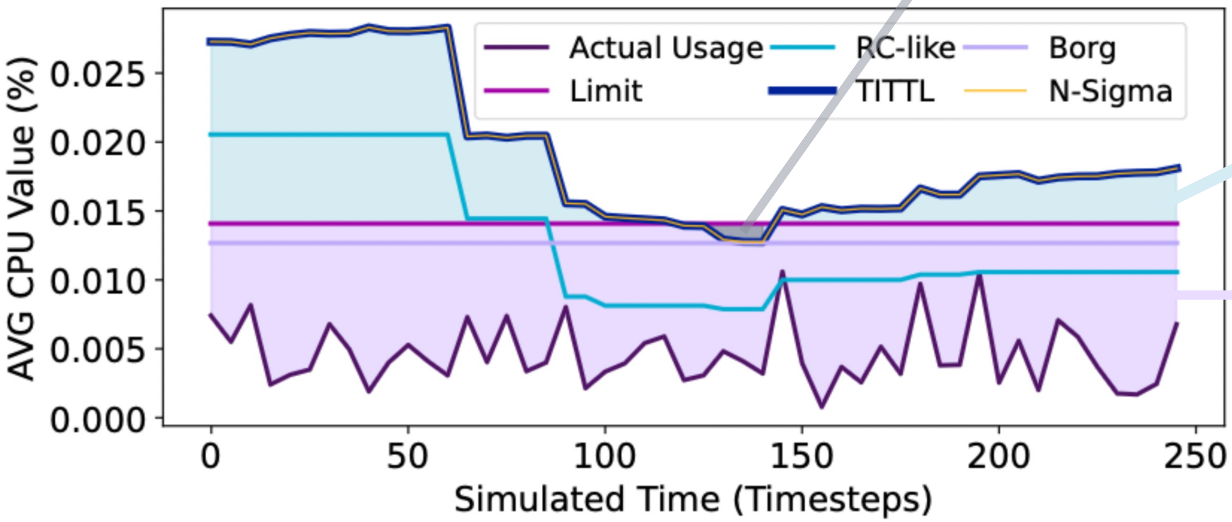
😬 **NO overcommitment** is happening for **94%** of the cases we examined, due to the predictor's **OVER-estimations.**

😱 Predictors just protect from under-estimations, ***allowing*** overcommitment **only 6%** of the times.

[EuroMLSys '24] *Do Predictors for Resource Overcommitment Even Predict?* Georgia Christofidi and Thaleia Dimitra Doudali.

Missed Opportunity for Resource Savings

Usable resources.
Resource savings **(2%)**



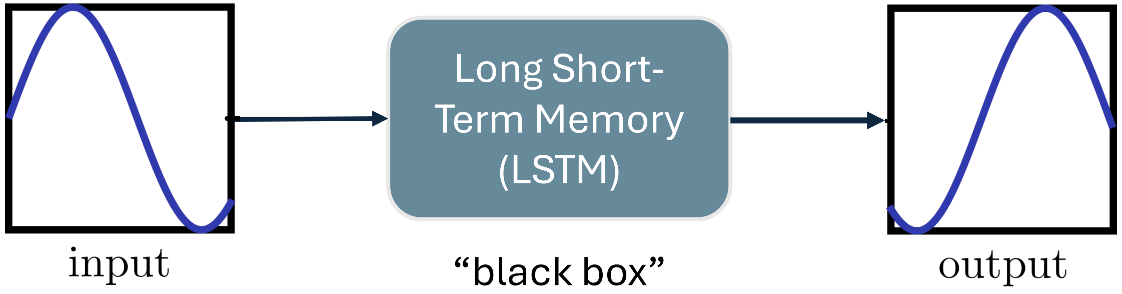
Completely wasted resources.

Maximum possible resource savings **(61%)** with an **accurate** predictor.

Here it seems to be ***necessary*** to have a more **accurate** and **intelligent** predictor!
At least a predictor that actually predicts!

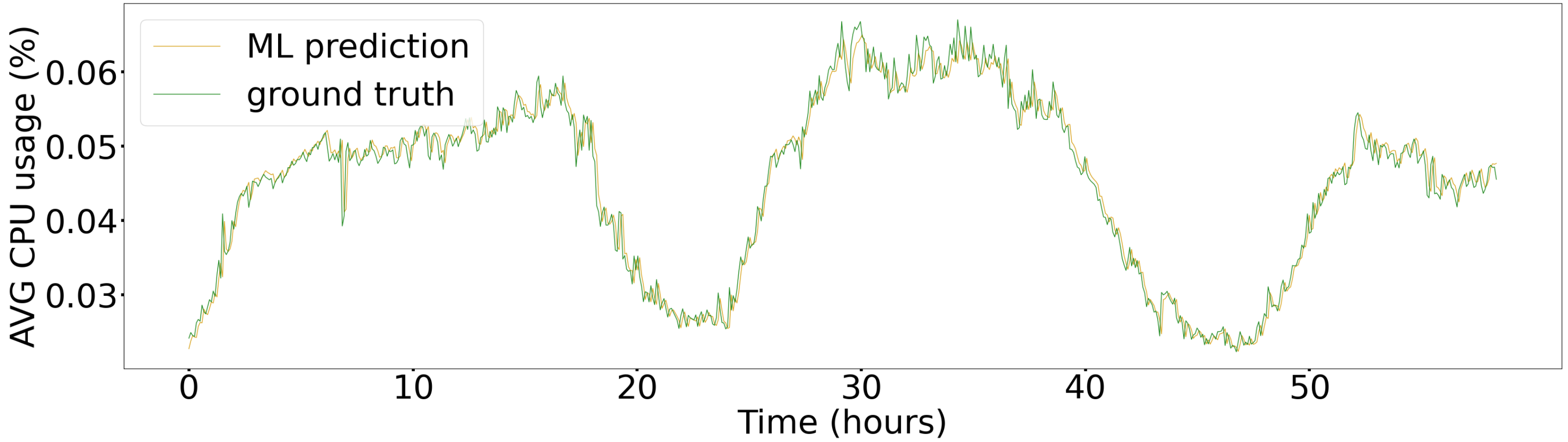
[EuroMLSys '24] *Do Predictors for Resource Overcommitment Even Predict?* Georgia Christofidi and Thaleia Dimitra Doudali.

Use Machine Learning? Let's use LSTMs!



Predict with high accuracy:

Traffic Conditions   Stock Market Prices   Weather

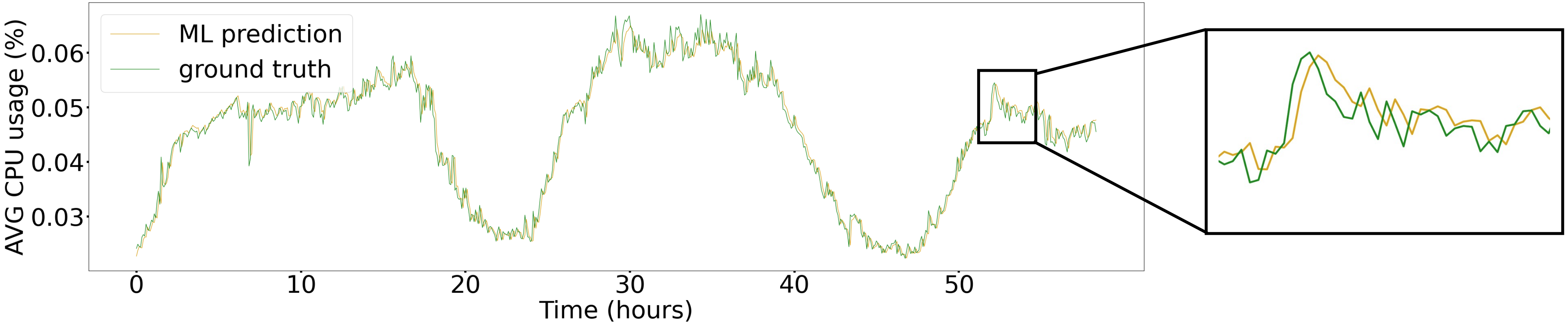


🥳 Looks quite accurate for cloud resource usage!!

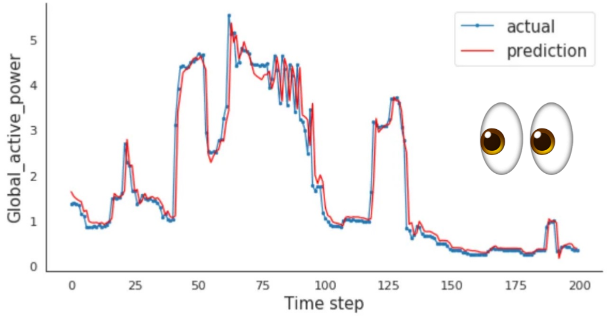
🥳 And seems to generalize well across patterns!

[EuroMLSys '23] Toward Pattern-based Model Selection for Cloud Resource Forecasting.
Georgia Christofidi , Konstantinos Papaioannou, Thaleia Dimitra Doudali.

Let's Take a Closer Look! 👀



Insight: LSTM predictions look like "shifted" versions of the real (ground truth) data.



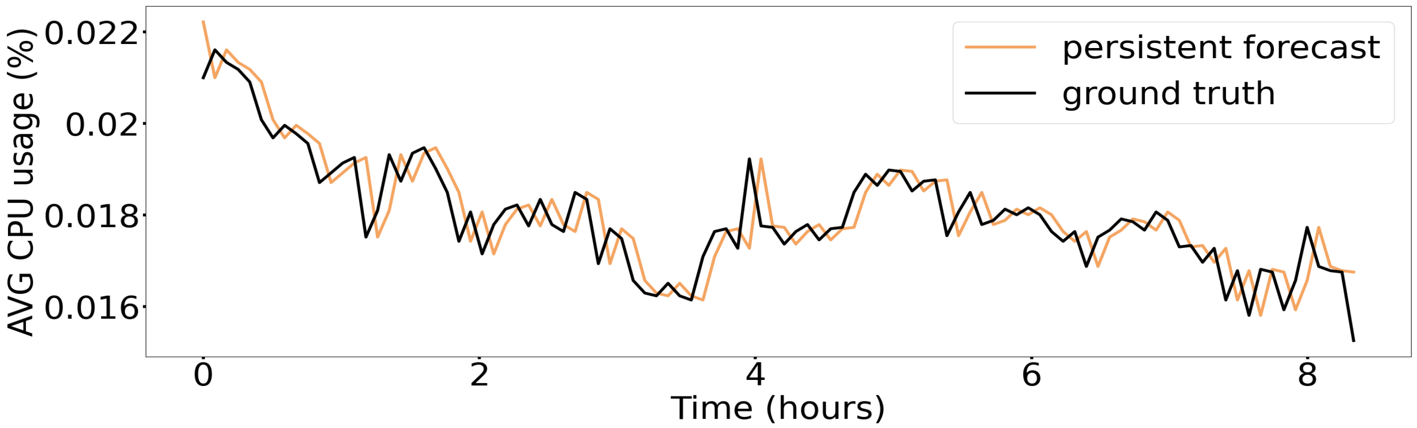
Source "Time Series Analysis, Visualization & Forecasting with LSTM" on towardsdatascience.com

<u>Lesson Learned</u>
Validate that ML *actually* learns!

[SoCC '23] *Is Machine Learning Necessary for Cloud Resource Usage Forecasting?*
Georgia Christofidi , Konstantinos Papaioannou, Thaleia Dimitra Doudali.

A Simple and Practical Predictor

Idea: Predict a shifted version of the ground truth, similar to the LSTMs.



Persistent Forecast**

Predicted Value(t) =

Ground Truth(t – 5 mins)

**[VLDB ‘21] Seagull: An Infrastructure for Load Prediction and Optimized Resource Allocation. By Microsoft.

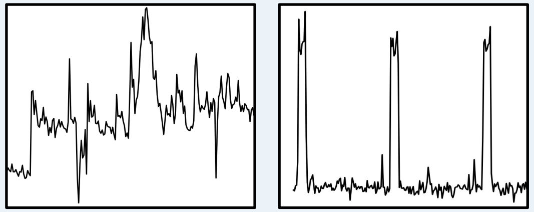
Simple, lightweight, explainable and easy to engineer in production-level.

Does it accurately predict resource usage?

[SoCC ‘23] *Is Machine Learning Necessary for Cloud Resource Usage Forecasting?*
Georgia Christofidi , Konstantinos Papaioannou, Thaleia Dimitra Doudali.

Extensive Experimental Evaluation

Public open-source datasets across different:

Cloud providers	Resource Types	Resource Levels	Usage Patterns	Frequency
Alibaba Cloud, Google Cloud, Microsoft Azure, Bitbrain	CPU, RAM	Physical Machine, Virtual Machine, Workload		Hourly/ Daily/ Weekly Windows

We calculate the **prediction error** of the persistent forecast.

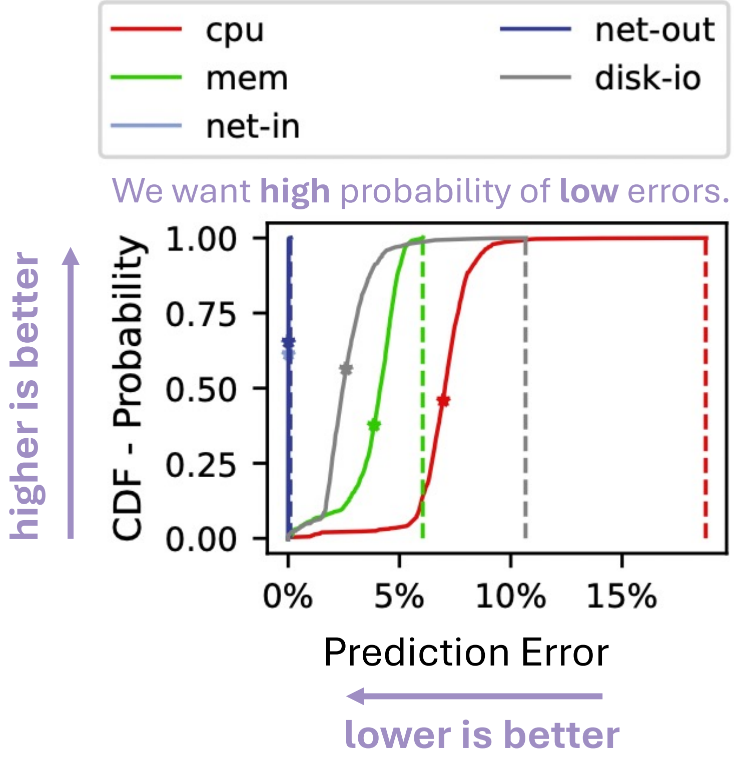
... will it work or do we *need* some other machine learning method?

[SoCC '23] *Is Machine Learning Necessary for Cloud Resource Usage Forecasting?*
Georgia Christofidi , Konstantinos Papaioannou, Thaleia Dimitra Doudali.

Results – Physical Machines

Physical Machine

Alibaba Dataset
Physical Machine Level



Observations:

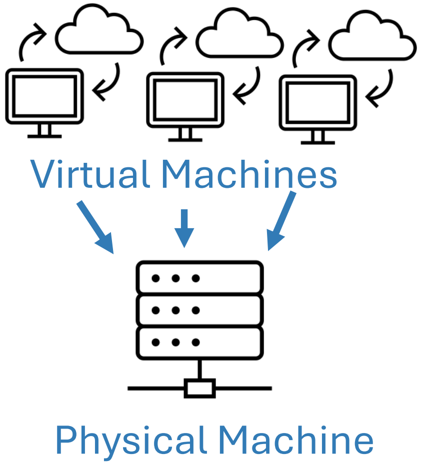
- net-in, net-out } Error ~ 0 🤩
- disk-io, mem } Avg. Error < 4 % Short tail. 😁
- cpu } Avg. Error ~ 7 % Longer tail. 😃

<u>Physical Machines</u>
Have **stable** load.

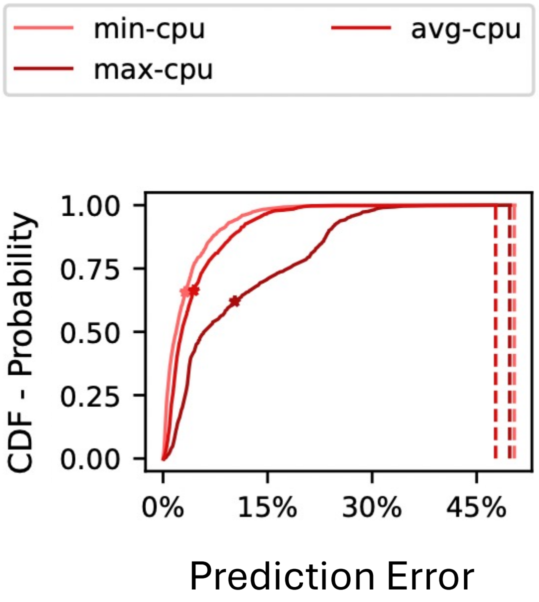
Persistent Forecast is **very** accurate!

[SoCC '23] *Is Machine Learning Necessary for Cloud Resource Usage Forecasting?*
Georgia Christofidi , Konstantinos Papaioannou, Thaleia Dimitra Doudali.

Results – Virtual Machine

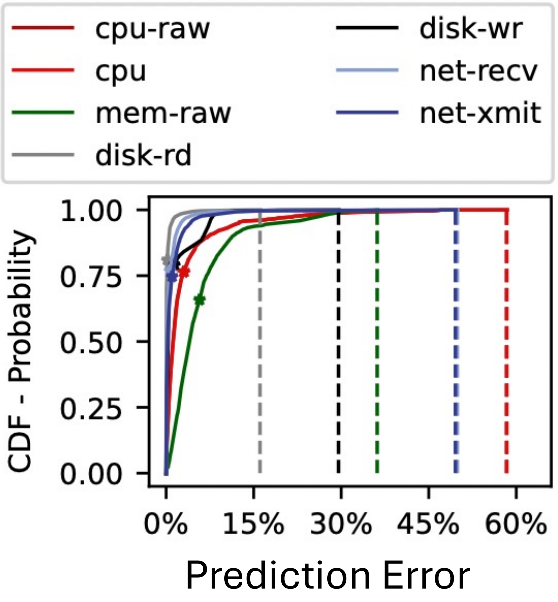


Azure Dataset
Virtual Machine Level

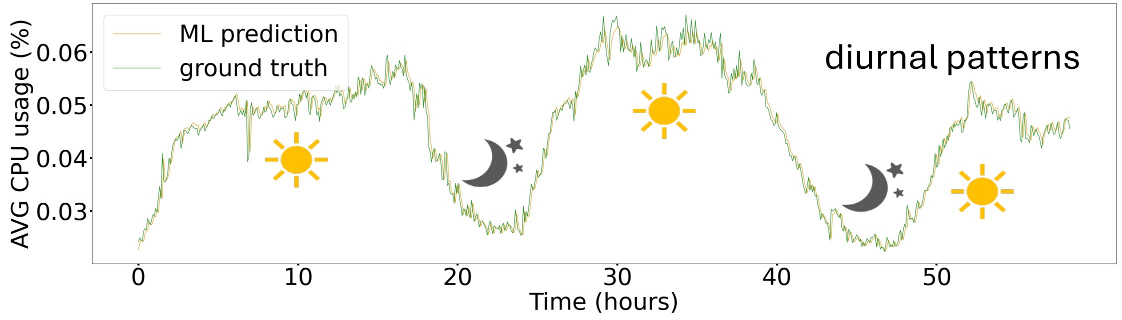


Average Prediction Error < 10%

Bitbrains Dataset
Virtual Machine Level



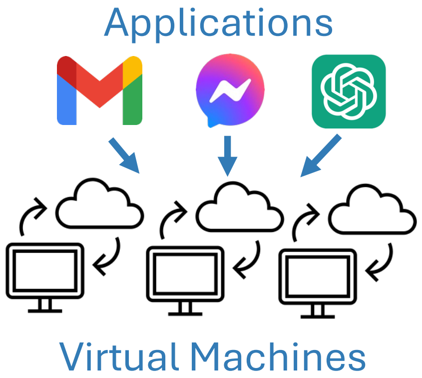
Average Prediction Error < 6 %



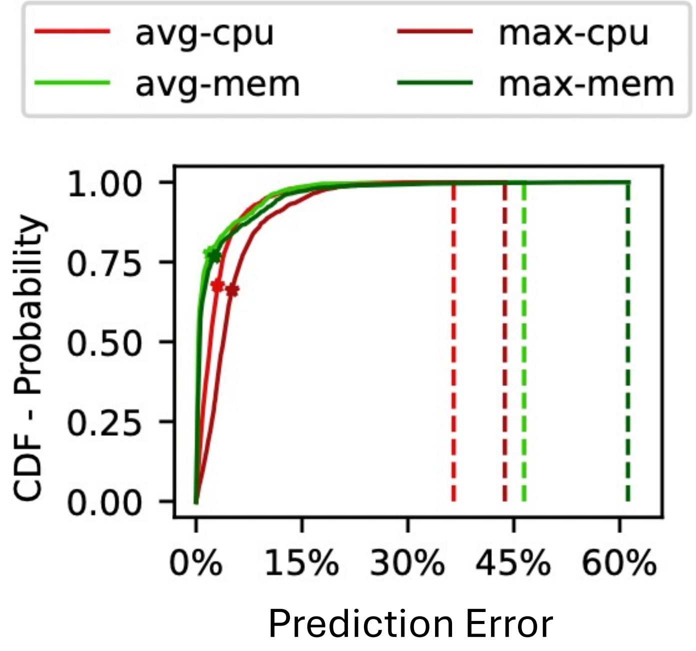
<u>Virtual Machines</u>
On average, **stable and periodic** load.
Patterns start becoming more dynamic.
(longer tails in the error)

[SoCC '23] *Is Machine Learning Necessary for Cloud Resource Usage Forecasting?*
Georgia Christofidi , Konstantinos Papaioannou, Thaleia Dimitra Doudali.

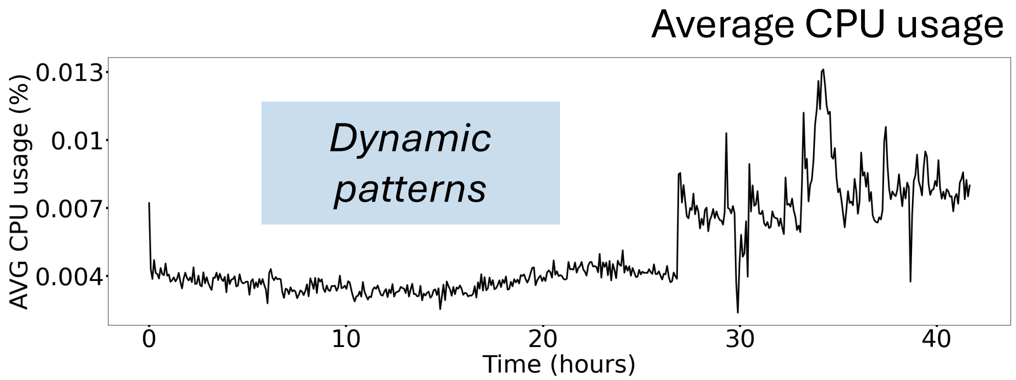
Results – Applications

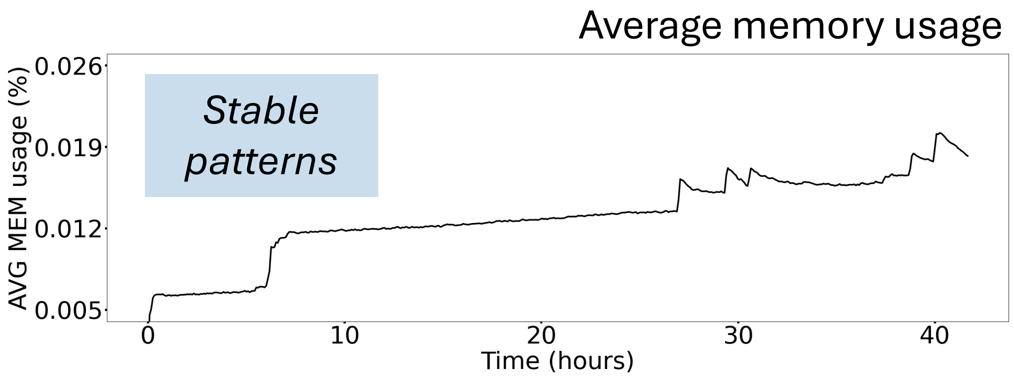


Google Dataset
Workload Level



Average Prediction Error < 6%





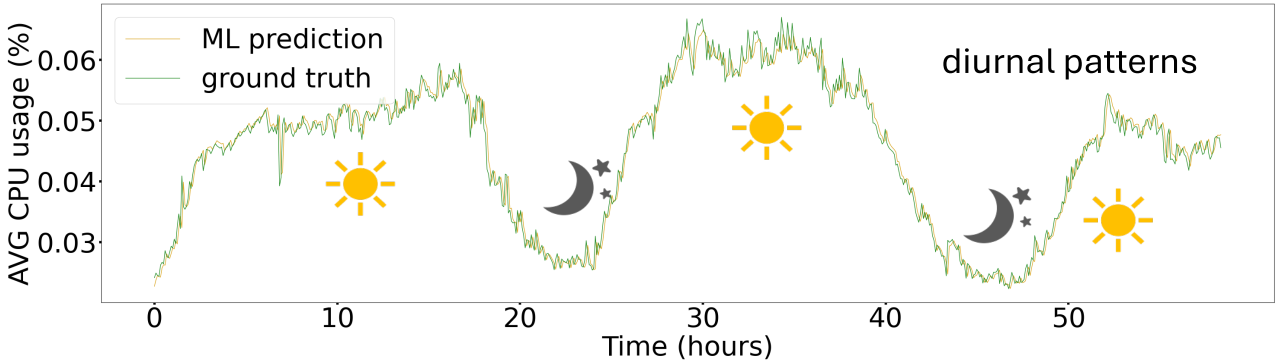
Applications

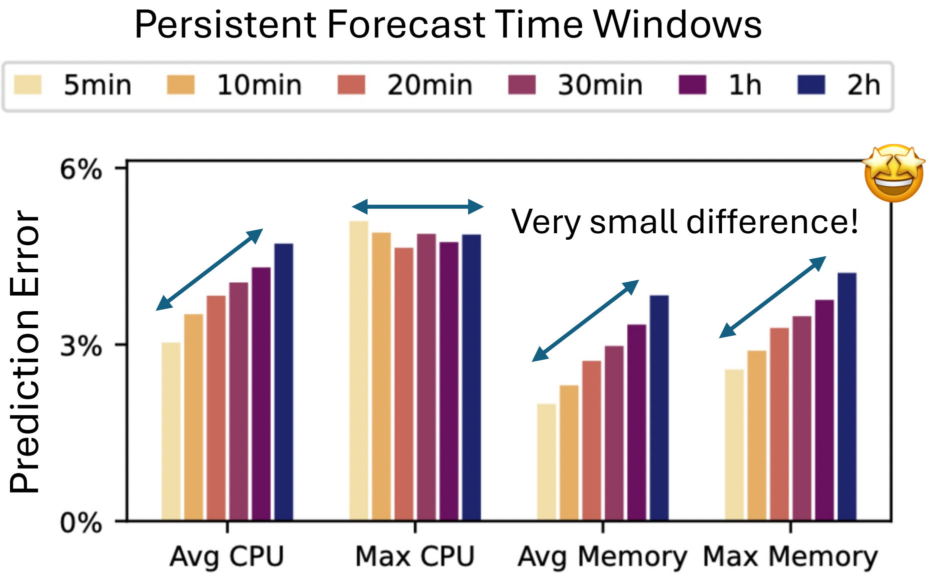
Most **dynamic** patterns.
(longest tails in the error)
Depends on the
type of resource!

[SoCC '23] *Is Machine Learning Necessary for Cloud Resource Usage Forecasting?*
Georgia Christofidi , Konstantinos Papaioannou, Thaleia Dimitra Doudali.

Why the Persistent Forecast Works?

Overall, on average, the persistent forecast is **very accurate**, prediction error < 6%. Why?



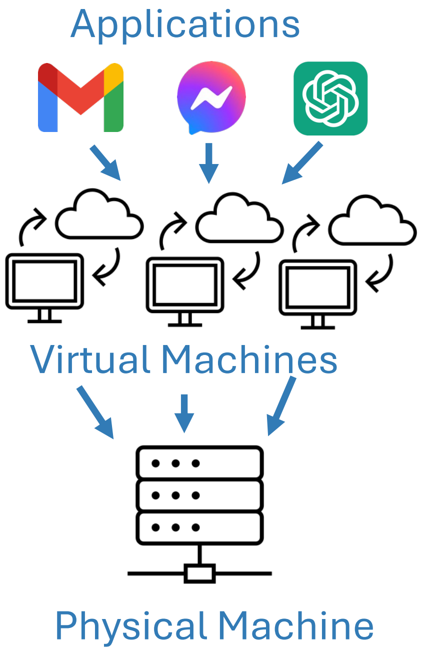


Because cloud resource usage is **highly persistent over time**, it changes very little every e.g. 5 minutes.

Low sensitivity to the time window length, reveals potential **repeating patterns** in the data.
This unlocks **opportunity** for even lower errors, if the time window matches the data periodicity.

[SoCC '23] *Is Machine Learning Necessary for Cloud Resource Usage Forecasting?*
Georgia Christofidi , Konstantinos Papaioannou, Thaleia Dimitra Doudali.

Is Machine Learning Necessary?



No!!

At least not always..

When is ML necessary?

Level	Pattern	ML?
Application	Dynamic	**Yes (?)**
Virtual Machine	Periodic	No
Physical Machine	Stable	No

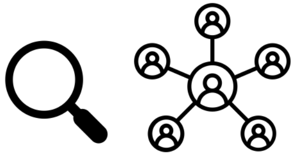
Type	Pattern	ML?
CPU	Dynamic	**Yes (?)**
Memory	Stable	No
Disk	Stable	No
Network	Stable	No

ML can be useful for reacting to unseen patterns: (1) if similar-seen: predict (2) if unseen: learn.

[SoCC ‘23] *Is Machine Learning Necessary for Cloud Resource Usage Forecasting?*
Georgia Christofidi , Konstantinos Papaioannou, Thaleia Dimitra Doudali.

New Questions and Challenges to Adress

1. When to Use ML?



Build data-driven methods to identify whether to use ML or not, e.g., Pattern detection and classification.

2. Which ML to use?

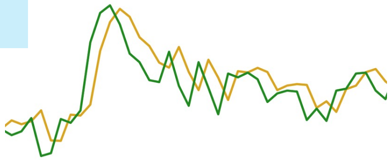
Probably not LSTMs.. 🫣 One of the current SOTA.. transformers? ChatGPT?

[AI4Sys @ HPDC 2024] *Toward Using Representation Learning for Cloud Resource Usage Forecasting*. Razine Ghorab, Thaleia Dimitra Doudali.

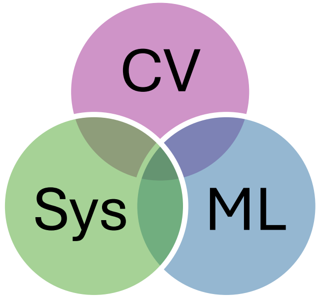
Production likes **simple and explainable**, e.g. decision trees?

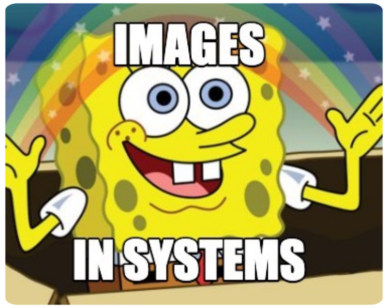
3. Validate ML actually learns!

Definitely use visual validation!



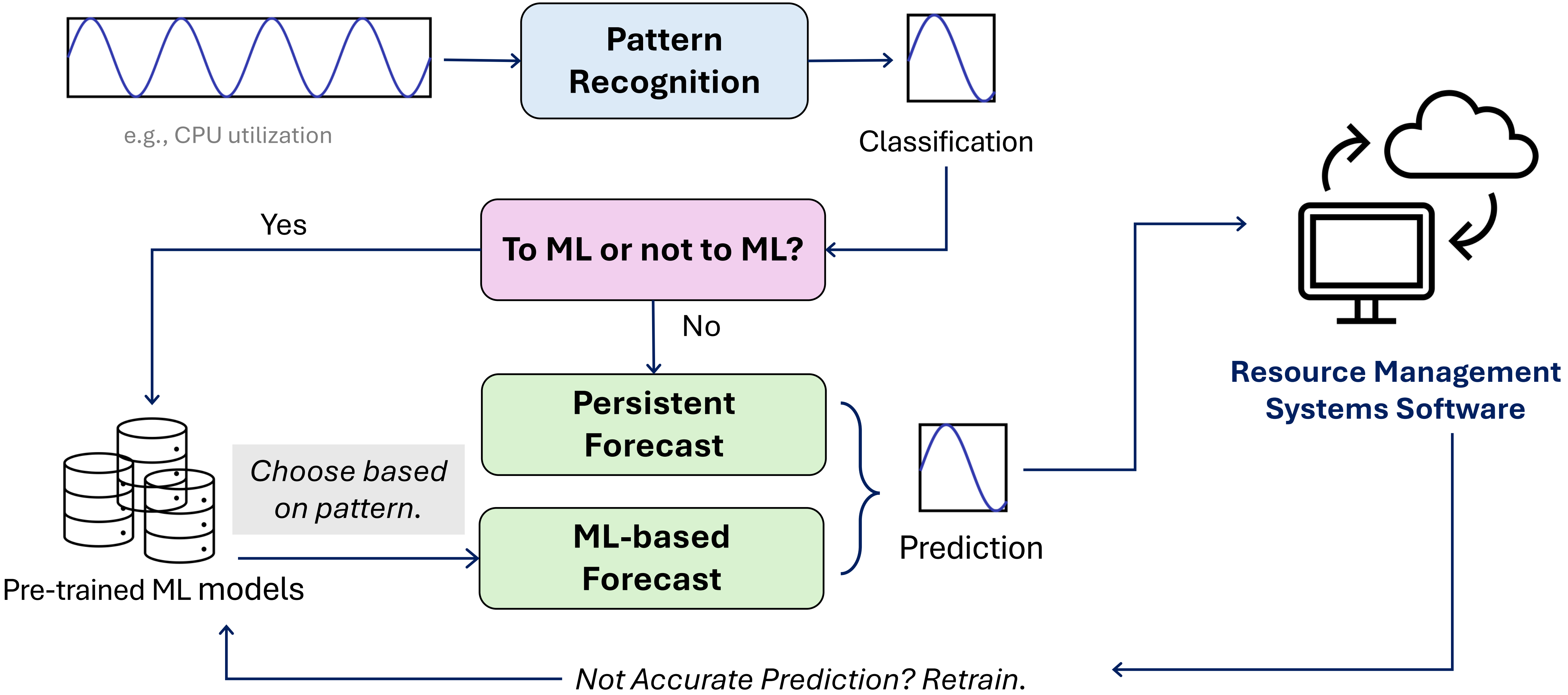
Maybe even some **Computer Vision (CV)** methods?





[Wild and Crazy Ideas @ ASPLOS 2022] A picture is worth a 1000… features!
Using Computer Vision alongside Machine Learning in Computer Systems Thaleia Dimitra Doudali.

Use ML to Augment, not Replace** - Our Vision

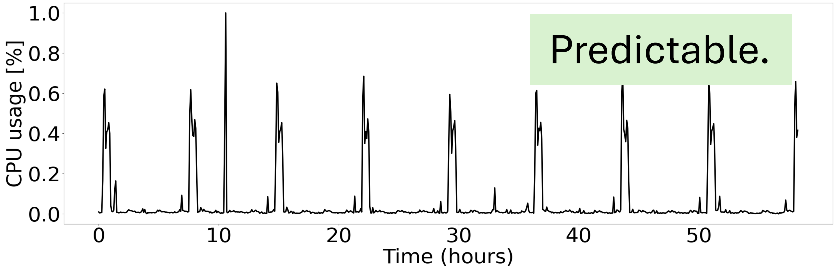


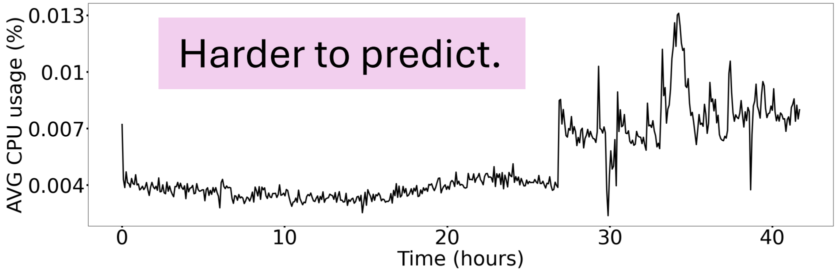
**[SIGMOD 2023 Companion] Towards Building Autonomous Data Services on Azure. (Microsoft agrees!)

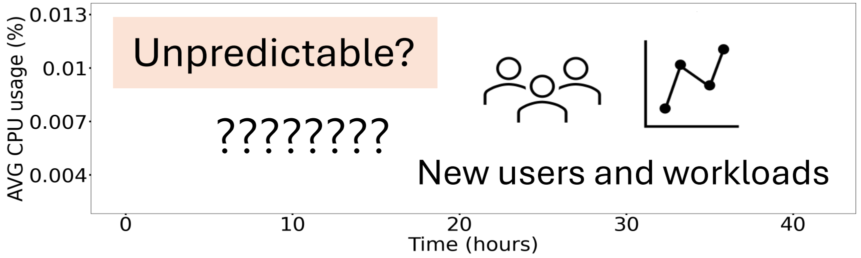
Summary

Machine Learning is *not always* necessary in cloud resource management.

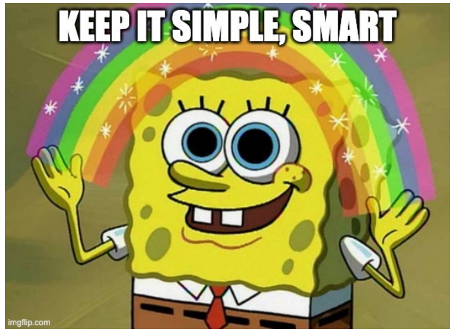
It all depends on the patterns.







Use **simple, explainable, lightweight** ML methods to **augment, not replace** robust analytical models, like the **persistent forecast.**



[SIGOPS Blog] KISS: Keep it Simple, Smart.
[SIGARCH Blog] Think Twice Before Using Machine Learning to Manage Cloud Resources.
Both written by Thaleia Dimitra Doudali

Thanks to my brilliant PhD students! 🫶

Georgia Christofidi

Konstantinos Papaioannou

Website

GitHub