Smart

Fast

Systems

Machine Learning

Computer Vision

Building Smart and Fast Systems using Machine Learning and Computer Vision.

Thaleia Dimitra Doudali

Assistant Research Professor @IMDEA Software Institute

About Me

Born and raised
in Greece.

2015

Undergrad in ECE at
NTUA, Athens, Greece.

2021

PhD in CS at
Georgia Tech, Atlanta, USA.

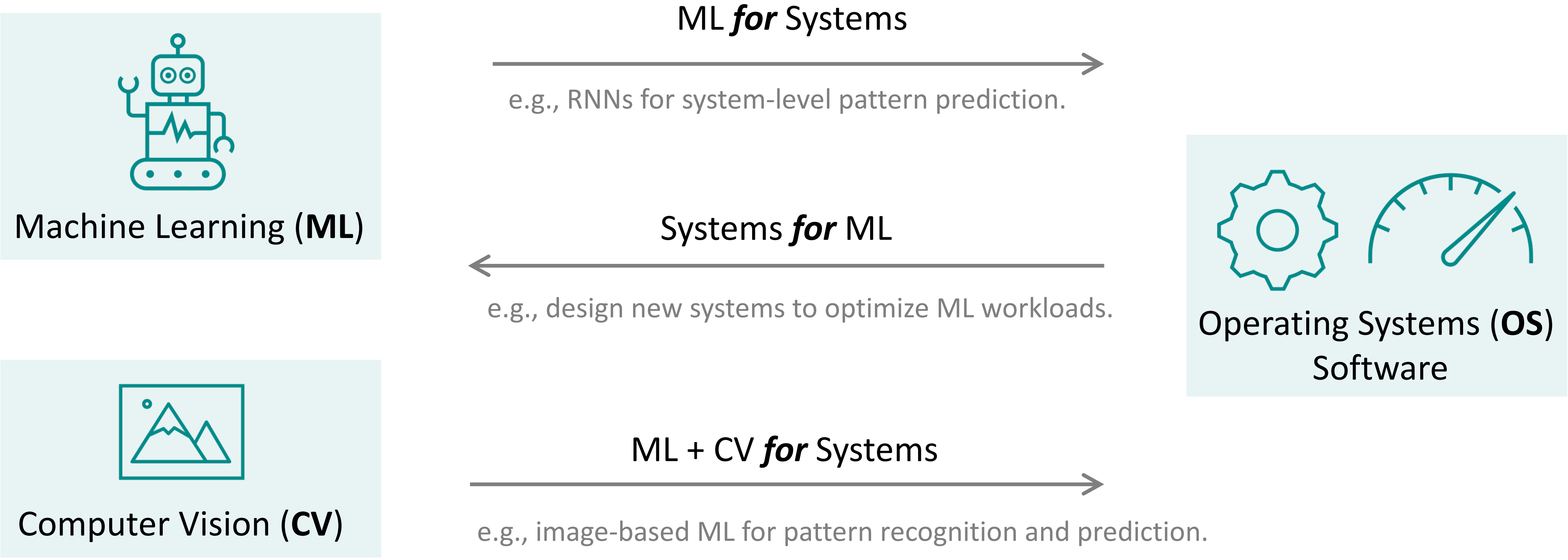
Advised by Ada Gavrilovska.

Start: October 2021

Assistant Professor at
IMDEA, Madrid, Spain.

About My Research

My research lies at the intersection of Machine Learning and Systems.



Machine Learning (**ML**)

Computer Vision (**CV**)

Operating Systems (**OS**) Software

ML ***for*** Systems

e.g., RNNs for system-level pattern prediction.

Systems ***for*** ML

e.g., design new systems to optimize ML workloads.

ML + CV ***for*** Systems

e.g., image-based ML for pattern recognition and prediction.

Talk Outline

Why do we need Smarter and Faster Systems?
The evolution of the hardware technologies, calls for software improvements.

Building *Smart* Systems
Using machine and human intelligence to build practical ML-based systems.

Building *Fast* Systems
Reducing ML-based management overheads with visualization.
Building image-based system pipelines.

Future Research Directions

Talk Outline

Why do we need Smarter and Faster Systems?
The evolution of the hardware technologies, calls for software improvements.

Building *Smart* Systems
Using machine and human intelligence to build practical ML-based systems.

Building *Fast* Systems
Reducing ML-based management overheads with visualization.
Building image-based system pipelines.

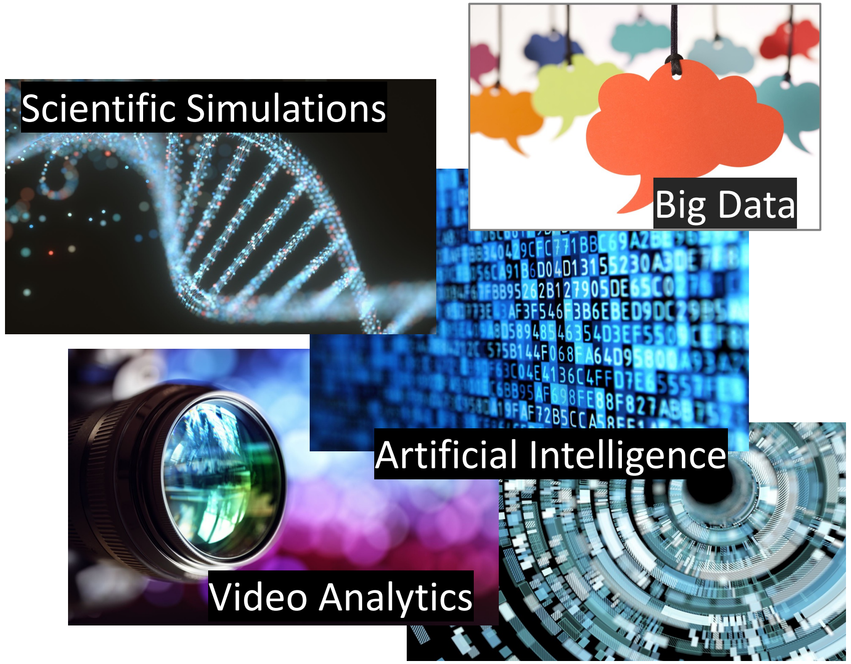
Future Research Directions

The Era of Data

“More than **65 ZB** of data will be created, captured, copied, and consumed in the world this year.”

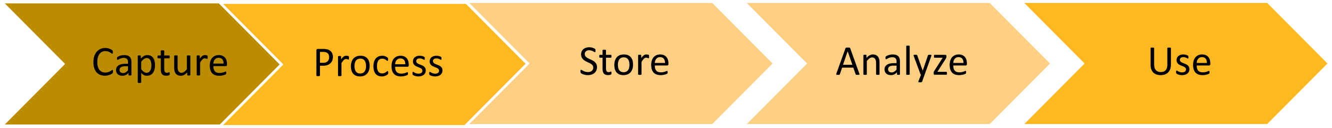
Source: International Data Corporation, March 2021.

Exploded Data Sizes



Data Analytics Pipeline

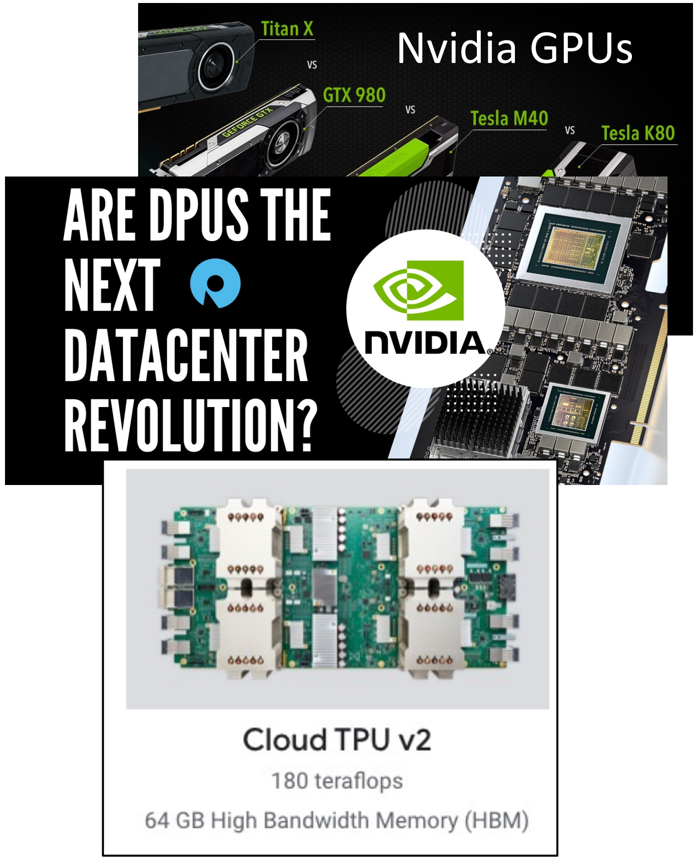
ZBs of data



Need for speed and massive storage capacities!

The Era of Heterogeneous Hardware

Compute Acceleration



Google

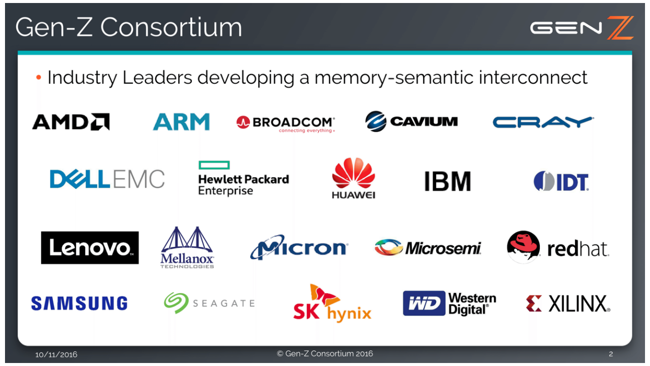
Data Storage Acceleration

Network Acceleration

Mellanox Innova™-2 Flex
Open Programmable SmartNIC

Interconnection Standards

CXL Compute Express Link™



Gen-Z Consortium

GEN Z

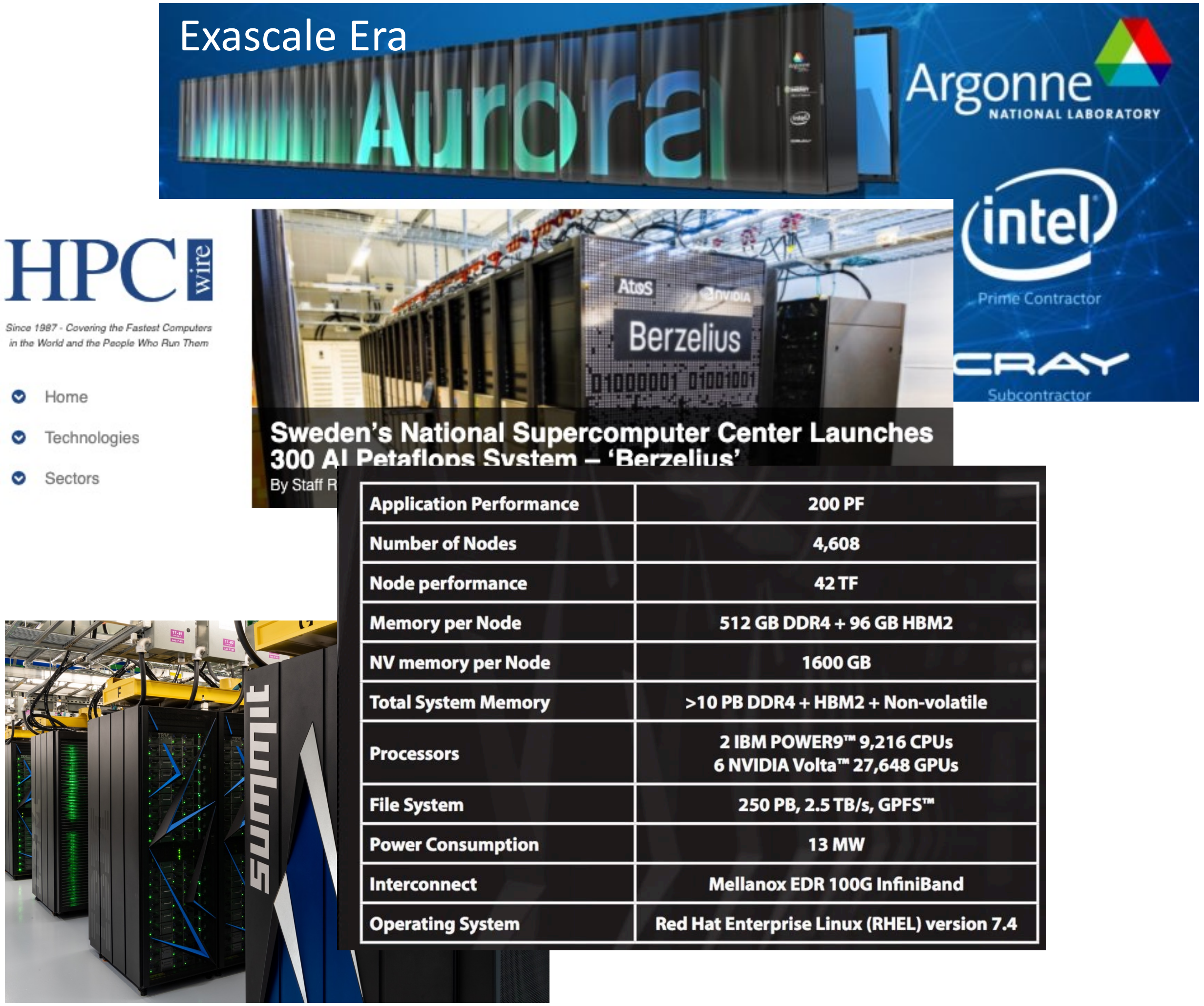
- Industry Leaders developing a memory-semantic interconnect

AMD ARM BROADCOM CAVIUM CRAY
DELL EMC Hewlett Packard Enterprise HUAWEI IBM IDT
Lenovo Mellanox Micron Microsemi redhat
SAMSUNG SEAGATE SK hynix Western Digital XILINX

10/11/2016 © Gen-Z Consortium 2016 2

Heterogeneity Across Computing Platforms

Supercomputers



Exascale Era

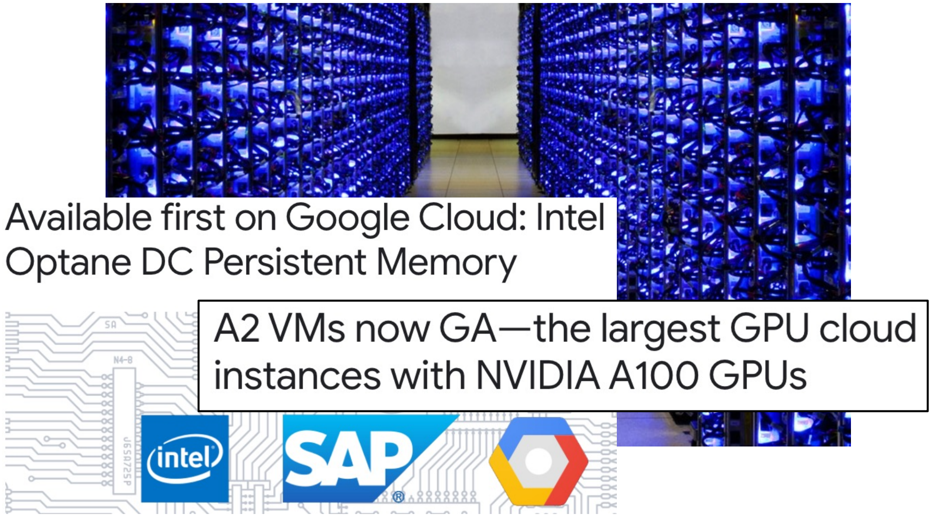
Aurora
Argonne NATIONAL LABORATORY
intel
Prime Contractor
CRAY
Subcontractor
HPC wire
Since 1987 - Covering the Fastest Computers in the World and the People Who Run Them
Home
Technologies
Sectors
Atos
NVIDIA
Berzelius
summit

Sweden's National Supercomputer Center Launches 300 AI Petaflops System – 'Berzelius'

By Staff R

Application Performance	200 PF
Number of Nodes	4,608
Node performance	42 TF
Memory per Node	512 GB DDR4 + 96 GB HBM2
NV memory per Node	1600 GB
Total System Memory	>10 PB DDR4 + HBM2 + Non-volatile
Processors	2 IBM POWER9™ 9,216 CPUs 6 NVIDIA Volta™ 27,648 GPUs
File System	250 PB, 2.5 TB/s, GPFS™
Power Consumption	13 MW
Interconnect	Mellanox EDR 100G InfiniBand
Operating System	Red Hat Enterprise Linux (RHEL) version 7.4

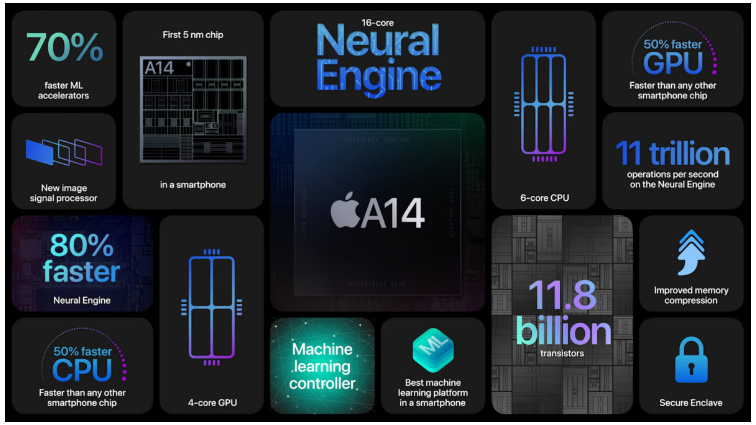
Datacenters



Available first on Google Cloud: Intel Optane DC Persistent Memory

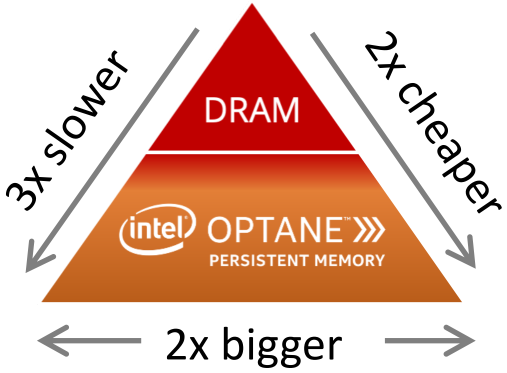
A2 VMs now GA—the largest GPU cloud instances with NVIDIA A100 GPUs

Personal Devices

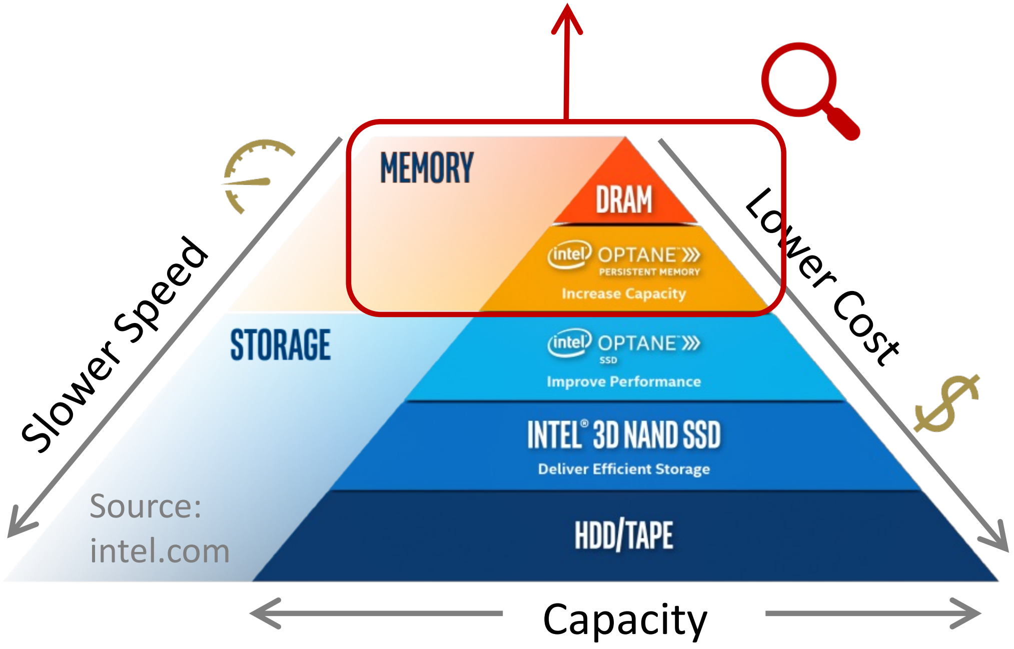


70%
faster ML accelerators
First 5 nm chip
A14
in a smartphone
16-core
Neural Engine
50% faster
GPU
Faster than any other smartphone chip
New image signal processor
A14
6-core CPU
11 trillion
operations per second on the Neural Engine
80% faster
Neural Engine
11.8 billion
transistors
Improved memory compression
50% faster
CPU
Faster than any other smartphone chip
4-core GPU
Machine learning controller
Best machine learning platform in a smartphone
Secure Enclave

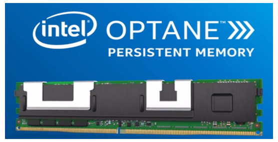
Heterogeneity Trade-offs



Source: memverge.com



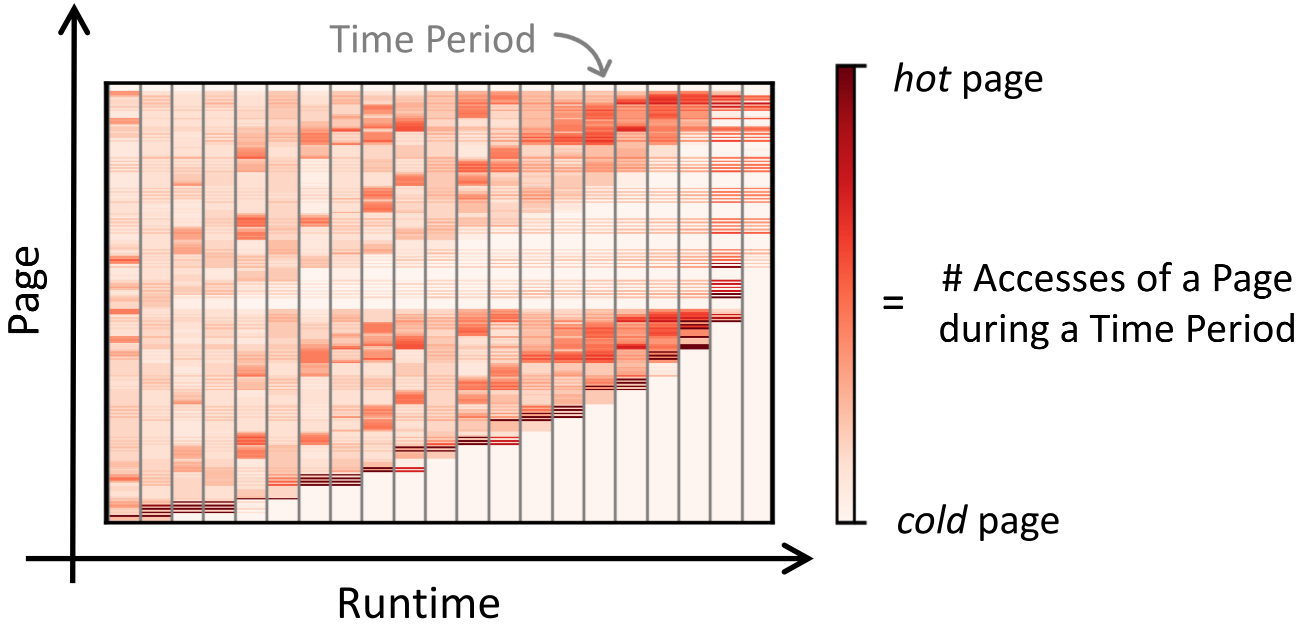
Source:
intel.com

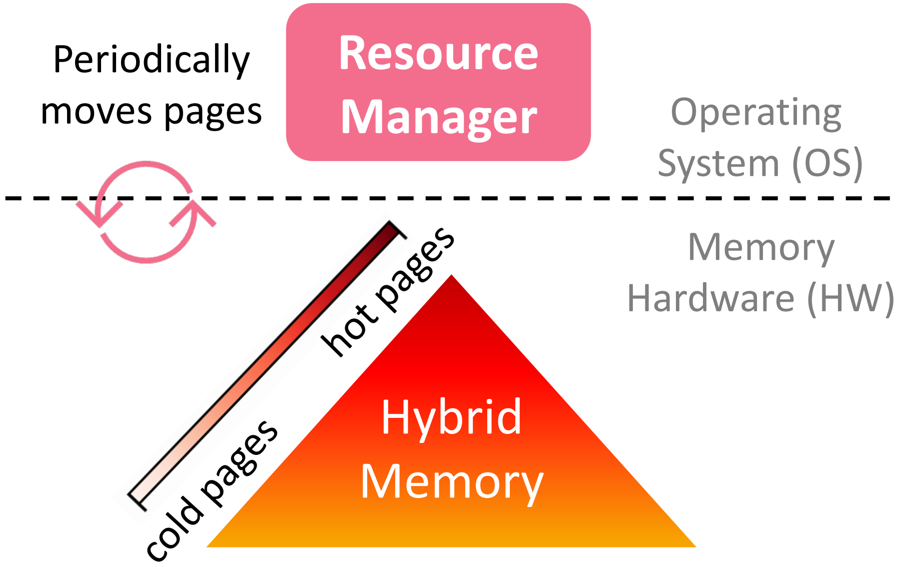
Characteristic	Technology	Vendors
Low Latency	MRAM	MRAM BY SPIN MEMORY EVERSPIN TECHNOLOGIES The MRAM Company Everspin Announces 1Gb ST-MRAM
Uniform Latency	DRAM	
High Bandwidth	HBM	
Persistent / Non Volatile	PMEM / NVM	

We are in the era of **Hybrid Memory** Systems.
A mix of different technologies at different speeds / capacities / costs.

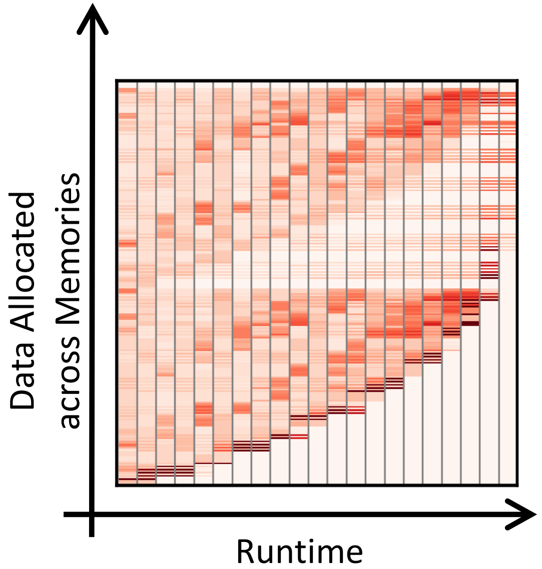
Hybrid Memory Management

The OS should move pages dynamically across hybrid memory to maximize the efficiency.





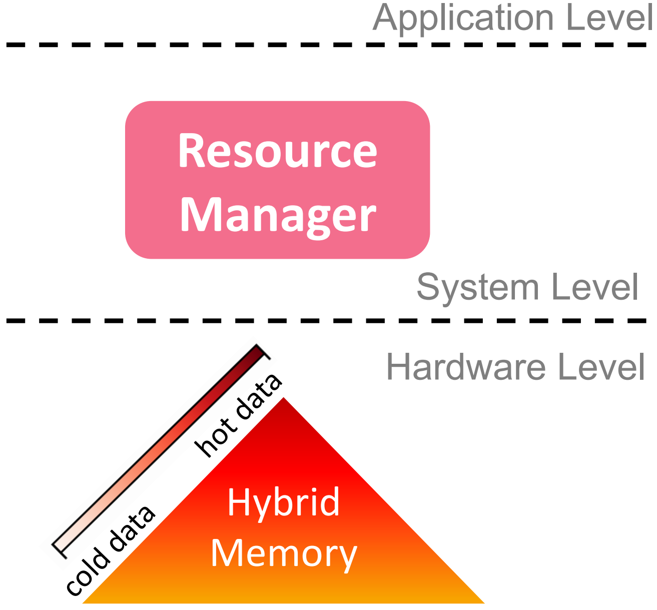
Need for Smarter Hybrid Memory Management



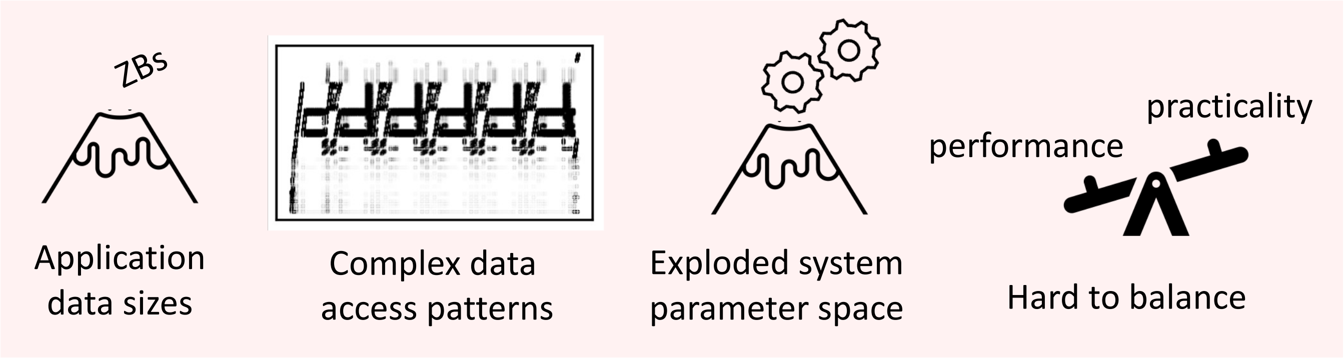
Dynamic Data Movements!

It is a **complex decision mix** to manage the data allocated across memories.

E.g., Which / How much / Where / When to move data?



Why do we need smarter and faster systems?



Application data sizes

Complex data access patterns

Exploded system parameter space

Hard to balance

Talk Outline

Why do we need Smarter and Faster Systems?
The evolution of the hardware technologies, calls for software improvements.

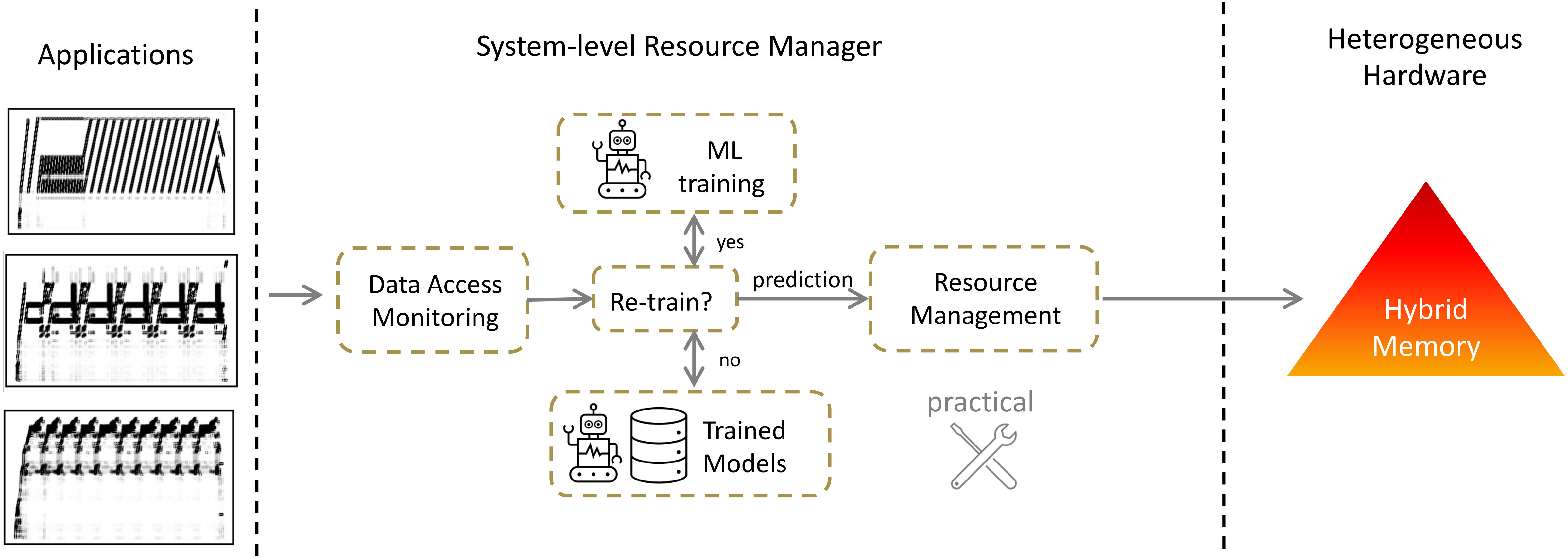
Building *Smart* Systems
Using machine and human intelligence to build practical ML-based systems.

Building *Fast* Systems
Reducing ML-based management overheads with visualization.
Building image-based system pipelines.

Future Research Directions

The Vision

ML-augmented heterogeneous resource manager.



Applications

System-level Resource Manager

Heterogeneous Hardware

ML training

yes

Data Access Monitoring

Re-train?

prediction

Resource Management

no

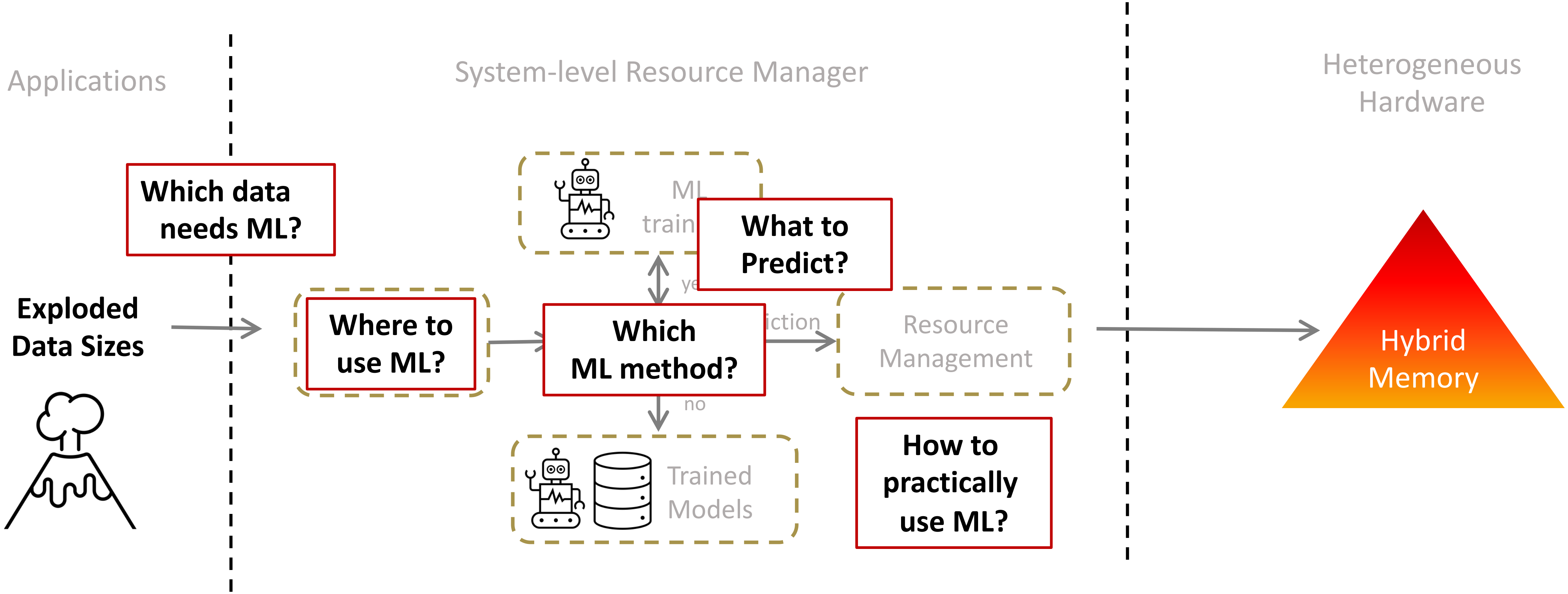
Trained Models

practical

Hybrid Memory

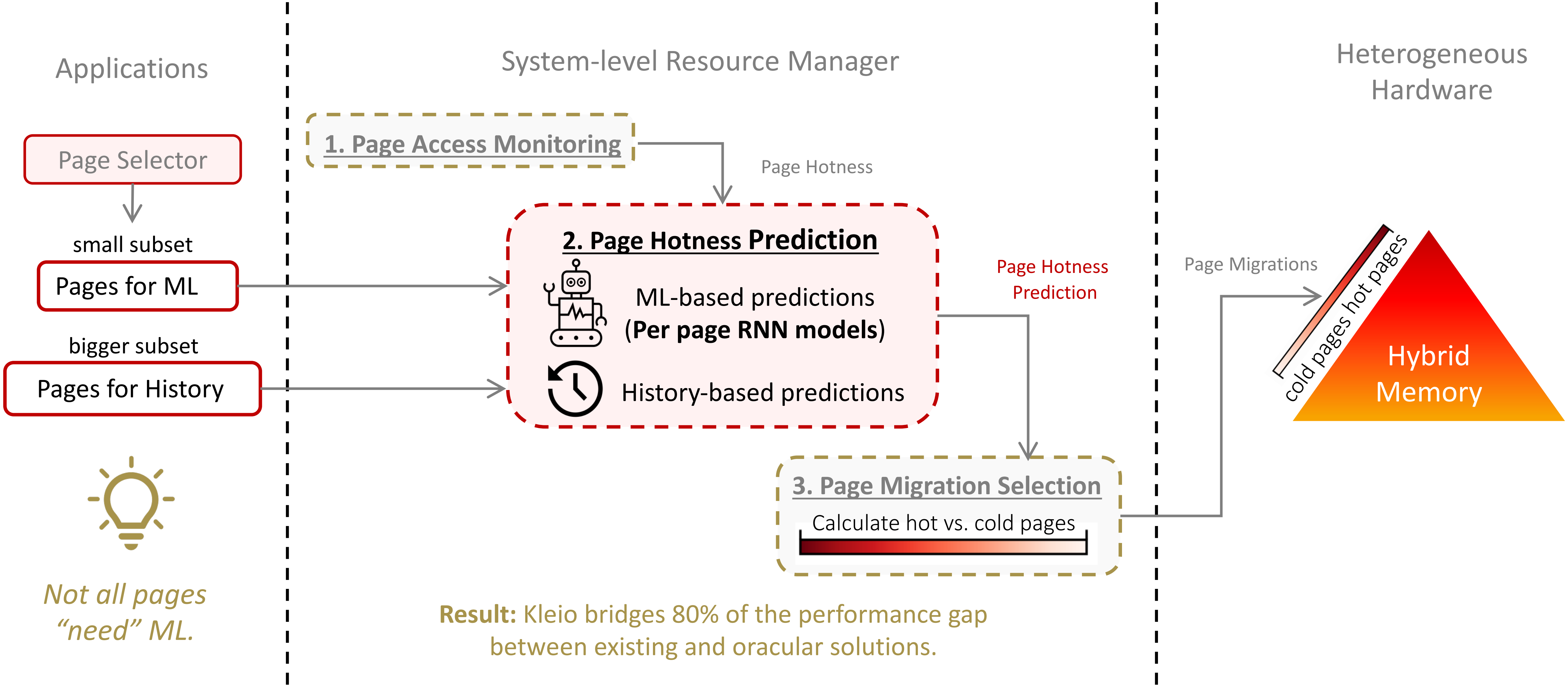
Contributions Towards the Vision

Laying the grounds for the *practical* integration of ML.



System design of **Kleio**

Kleio: a hybrid memory page scheduler with machine intelligence. [HPDC 2019]

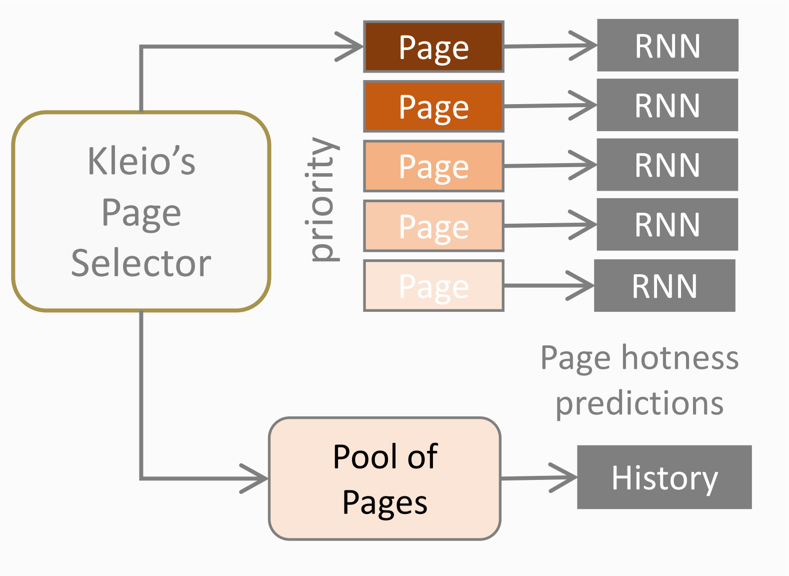


Not all pages "need" ML.

Result: Kleio bridges 80% of the performance gap between existing and oracular solutions.

The Key(s) to a Practical and Efficient ML-based System Design

Apply ML **when** and **where** necessary.

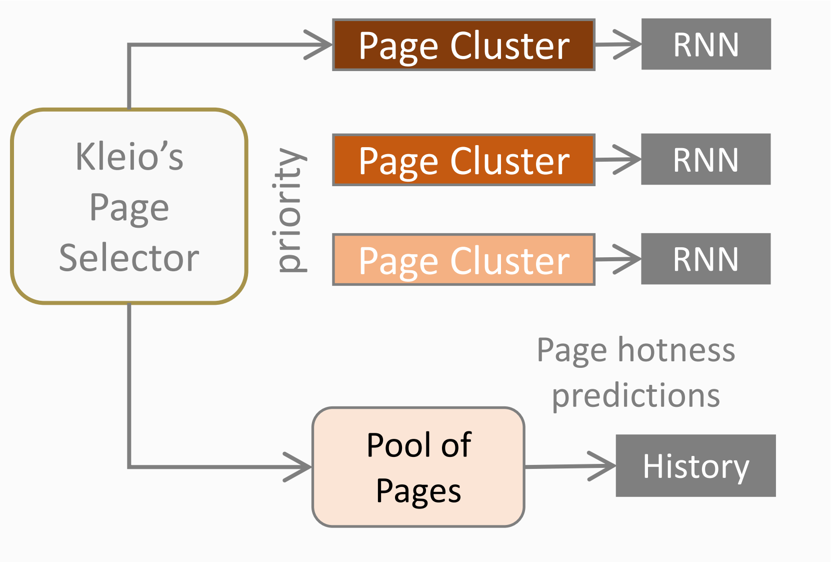


Apply ML on a small page subset.

Foundations for practical use of ML.

Carefully select pages for ML.

Application performance boost.

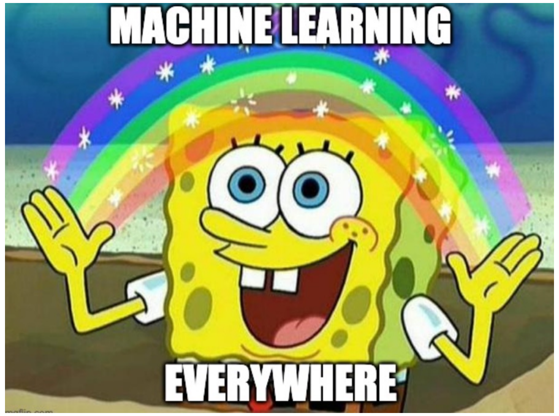


Small can still mean thousands of pages, because of the massive memory footprints of modern workloads.

Can we reduce the number of pages via clustering?

Insights from the System Design of **Coeus**

Coeus: Clustering (A)like Patterns for Practical Machine Intelligent Hybrid Memory Management . [CCGrid 2022]



Clustering? Let's use ML!

For example, K-means.

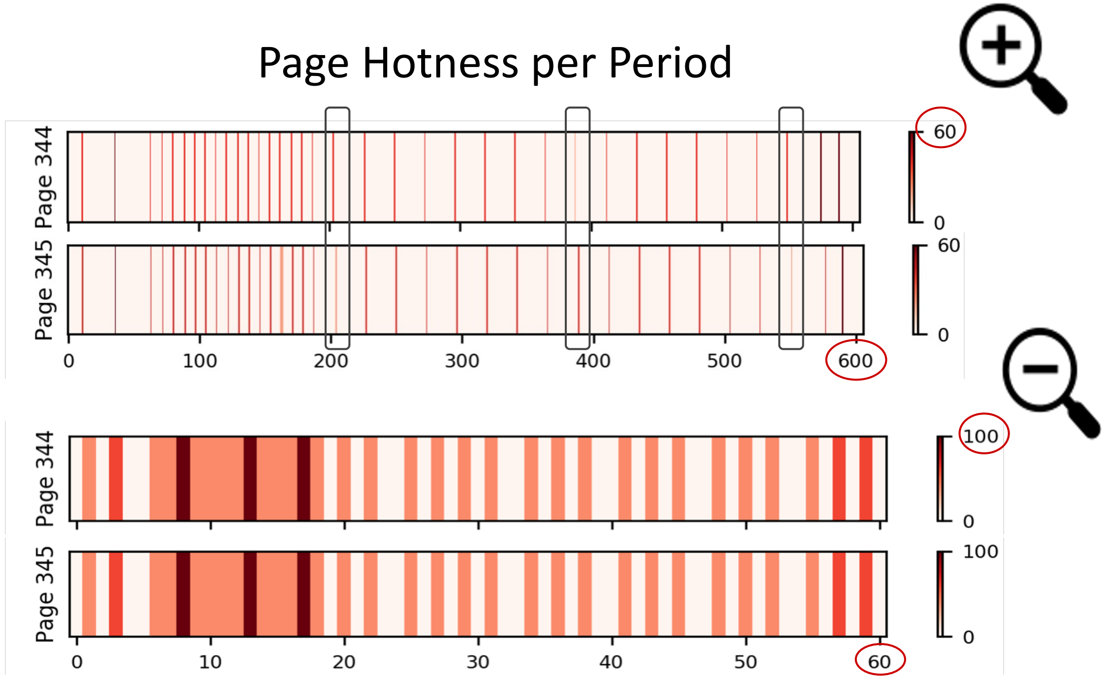
- How many clusters?
- Clustered input to ML?

Not trivial to configure.

Let's use our human intelligence..

.. Kleio learns the patterns of page hotness across time periods.

Page Hotness per Period



So what if I increase the duration of the period?

Key Idea

Group pages with *identical* patterns under a single ML model.

↓ 3x less RNNs

↑ 3x more performance

Talk Outline

Why do we need Smarter and Faster Systems?
The evolution of the hardware technologies, calls for software improvements.

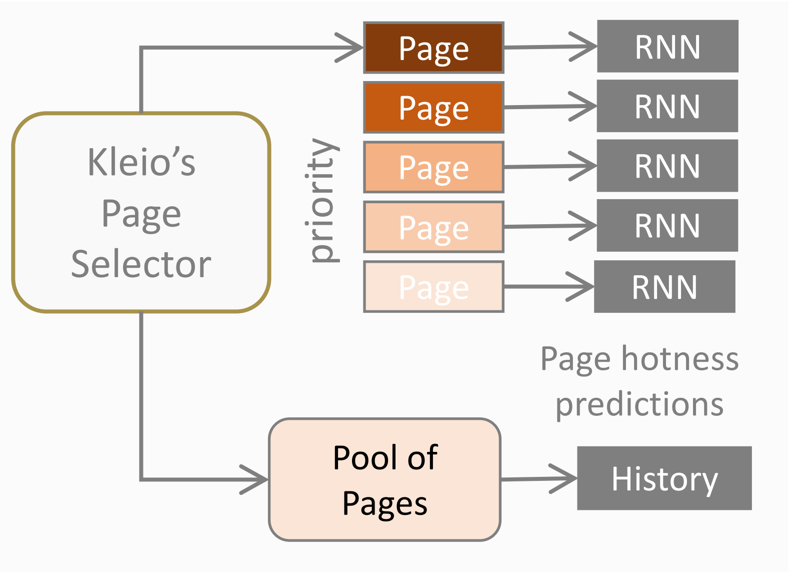
Building *Smart* Systems
Using machine and human intelligence to build practical ML-based systems.

Building *Fast* Systems
Reducing ML-based management overheads with visualization.
Building image-based system pipelines.

Future Research Directions

The Key(s) to a Practical and Efficient ML-based System Design

Apply ML **when** and **where** necessary.

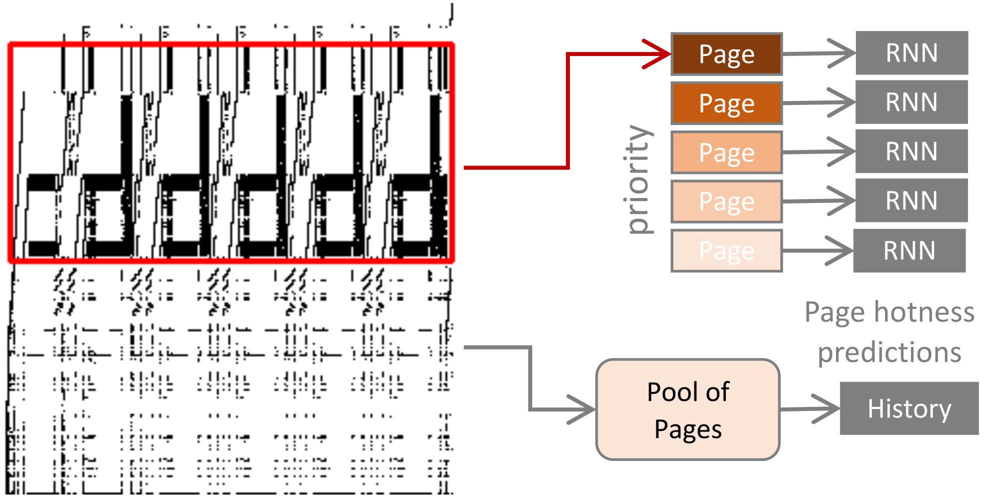


Apply ML on a small page subset.

Foundations for practical use of ML.

Carefully select pages for ML.

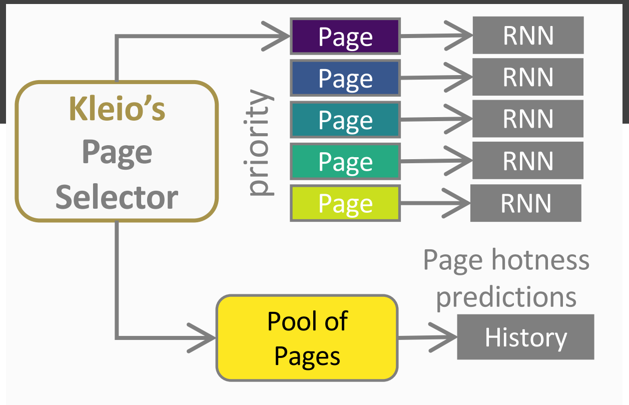
Application performance boost.

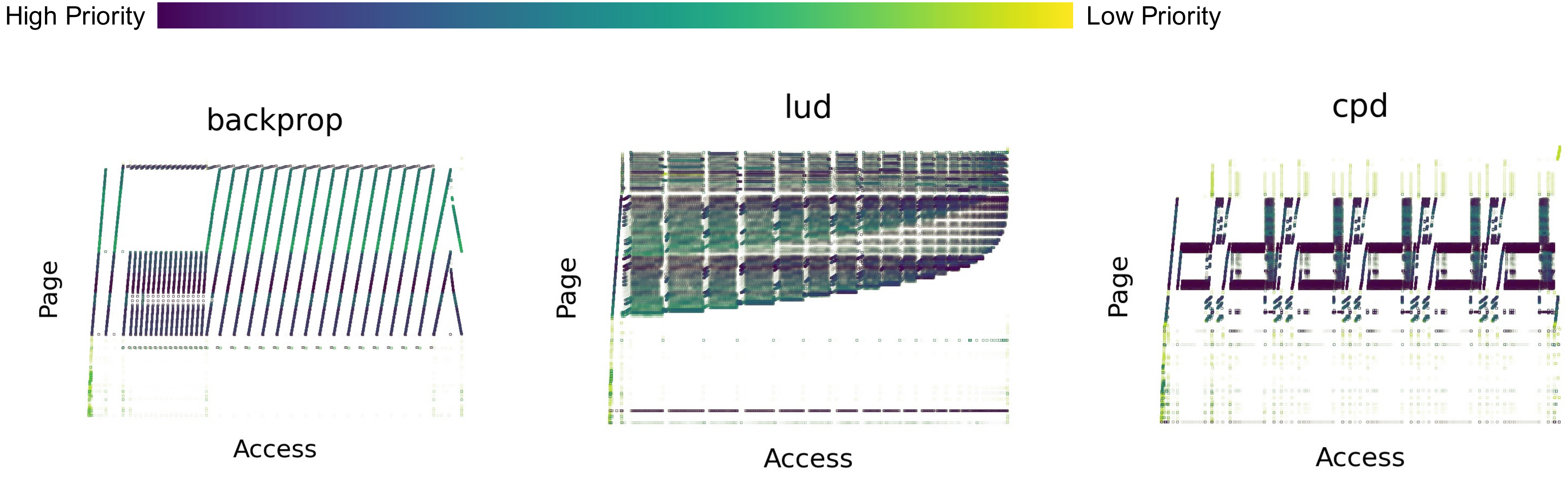


The page selection is not a lightweight process.
Performance modeling and estimations are used to maximize the effects of ML on application performance.

Can we accelerate the page selection process?

Insight from Visualizing Pages Selected for ML

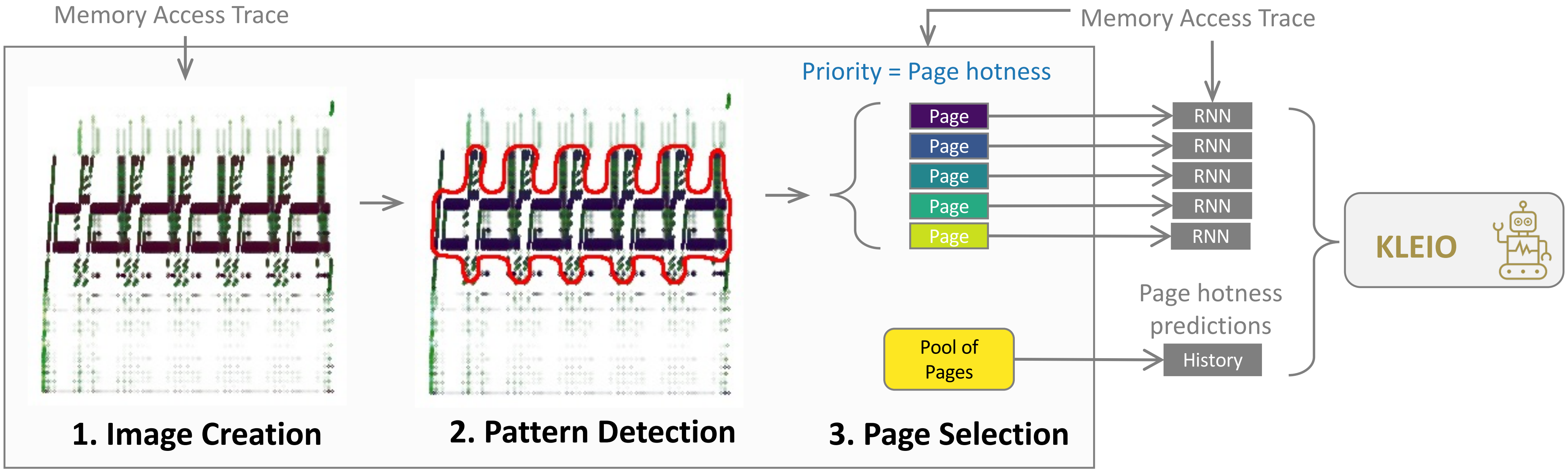




Neighboring pages that are part of distinct access patterns across *time* receive similar priority for ML.

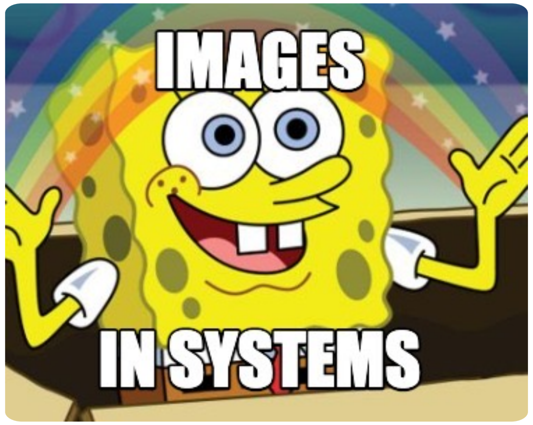
Towards Image-based Page Selection

Cronus: Computer Vision-based Machine Intelligent Hybrid Memory Management. [MEMSYS 2022]



Cronus reduces by **400x** the page selection times, from minutes down to seconds.

Why Use Images Inside Operating Systems?



Creating images helps:

- Another way to represent data, reducing their dimensionality to a 2D / 3D space.
- Captures spatial and temporal correlations.
- Leverage computer vision and image-based algorithms.

Feature Extraction

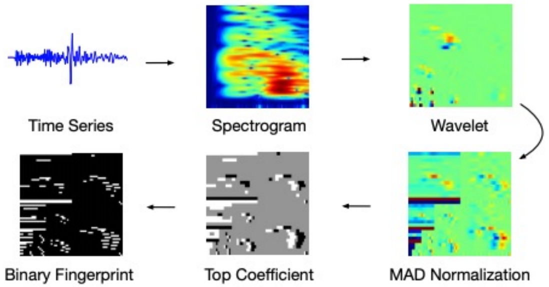


Figure 3: The fingerprinting algorithm encodes time-frequency features of the original time series into binary vectors.

Source: Kexin Rong et al. at VLDB ‘18.

Earthquake Detection:

Extract Frequencies of Seismic Waves.

Image-based ML Classifiers

Figure 1: Typical workstation of a professional trader. Credit: Photoagriculture / Shutterstock.com.

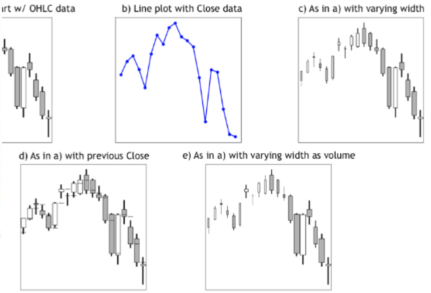
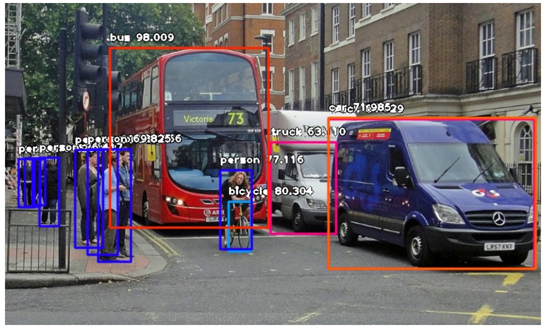


Figure 4: Various visual representations of the same time-series data.

Source:
J.P. Morgan AI labs.

Stock Market Forecasting:

Trading by learning time series data as images.



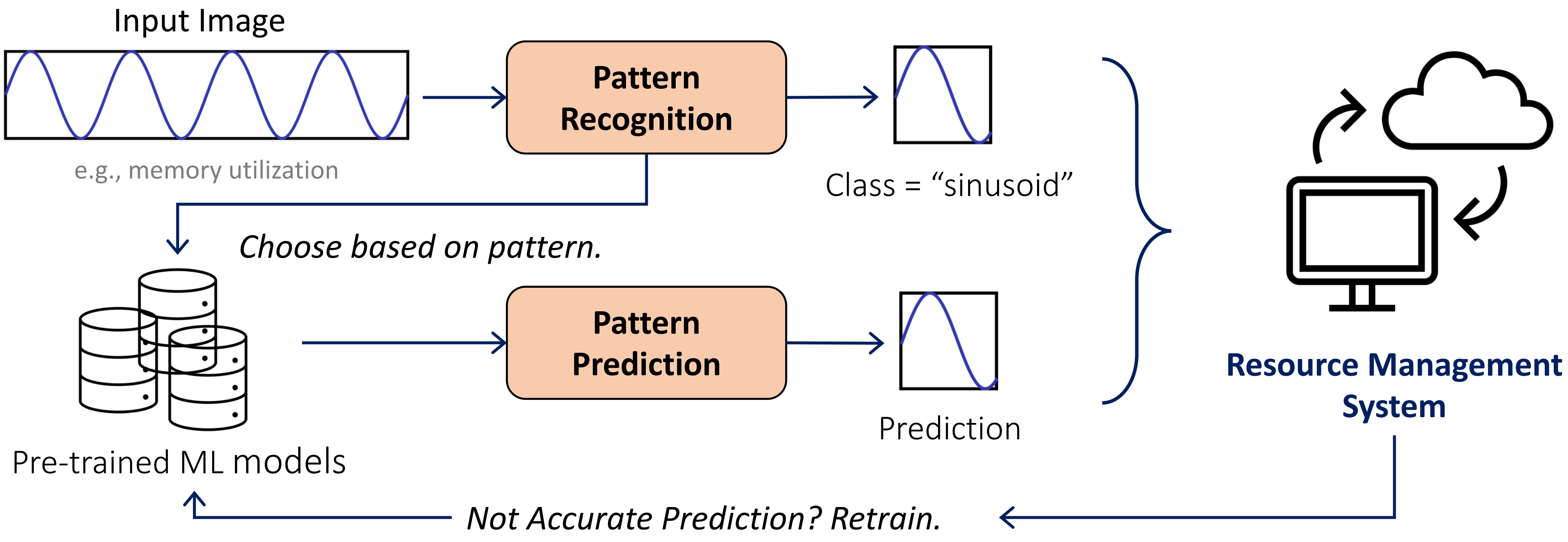
Autonomous Driving:

Object Detection & Recognition

Computer Vision + Machine Learning for Systems (1)

What can an image-based system pipeline look like?

E.g., predicting future resource utilization.



Input Image

e.g., memory utilization

Pattern Recognition

Class = "sinusoid"

Choose based on pattern.

Pre-trained ML models

Pattern Prediction

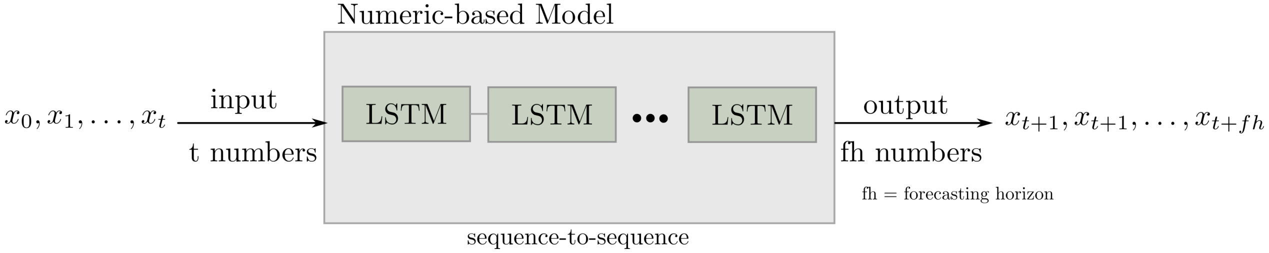
Prediction

Resource Management System

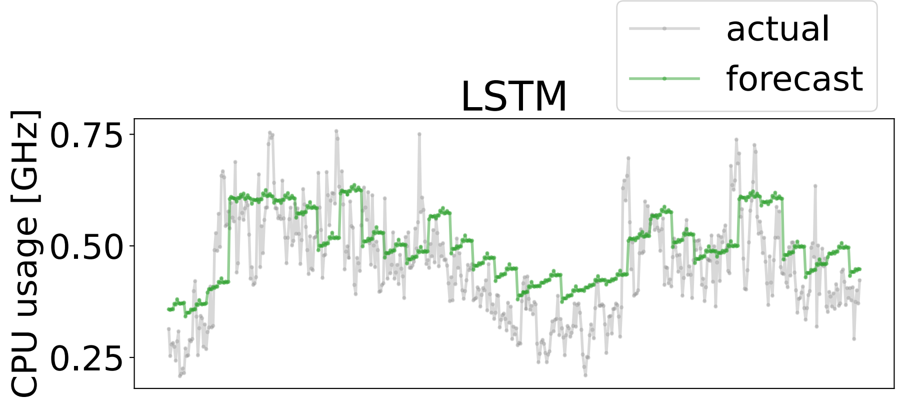
Not Accurate Prediction? Retrain.

Image-based vs. Number-based Machine Learning

Research paper under submission.



Number-based LSTM model



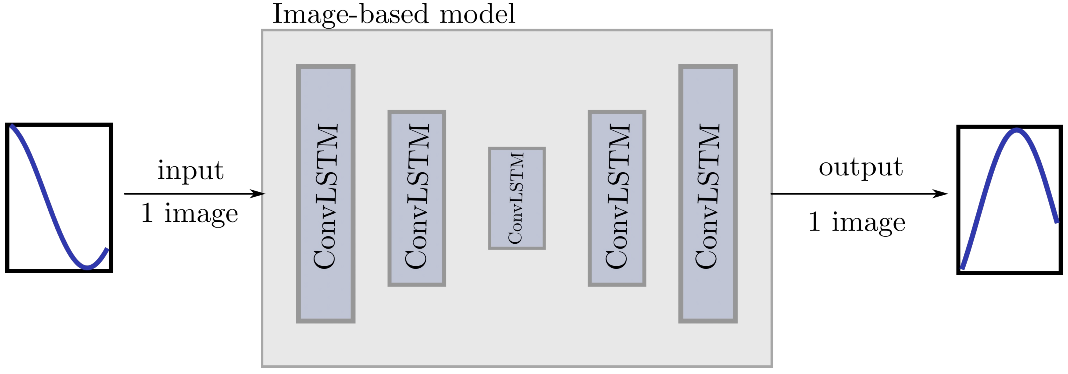
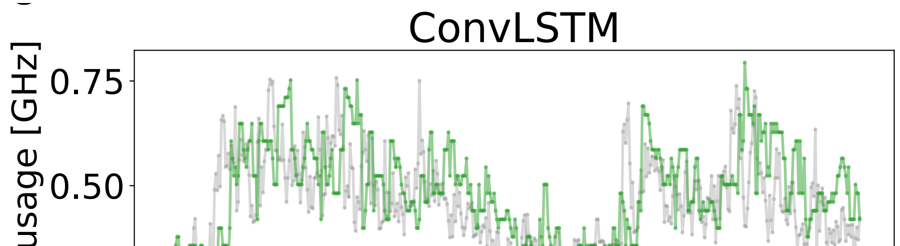


Image-based ConvLSTM model



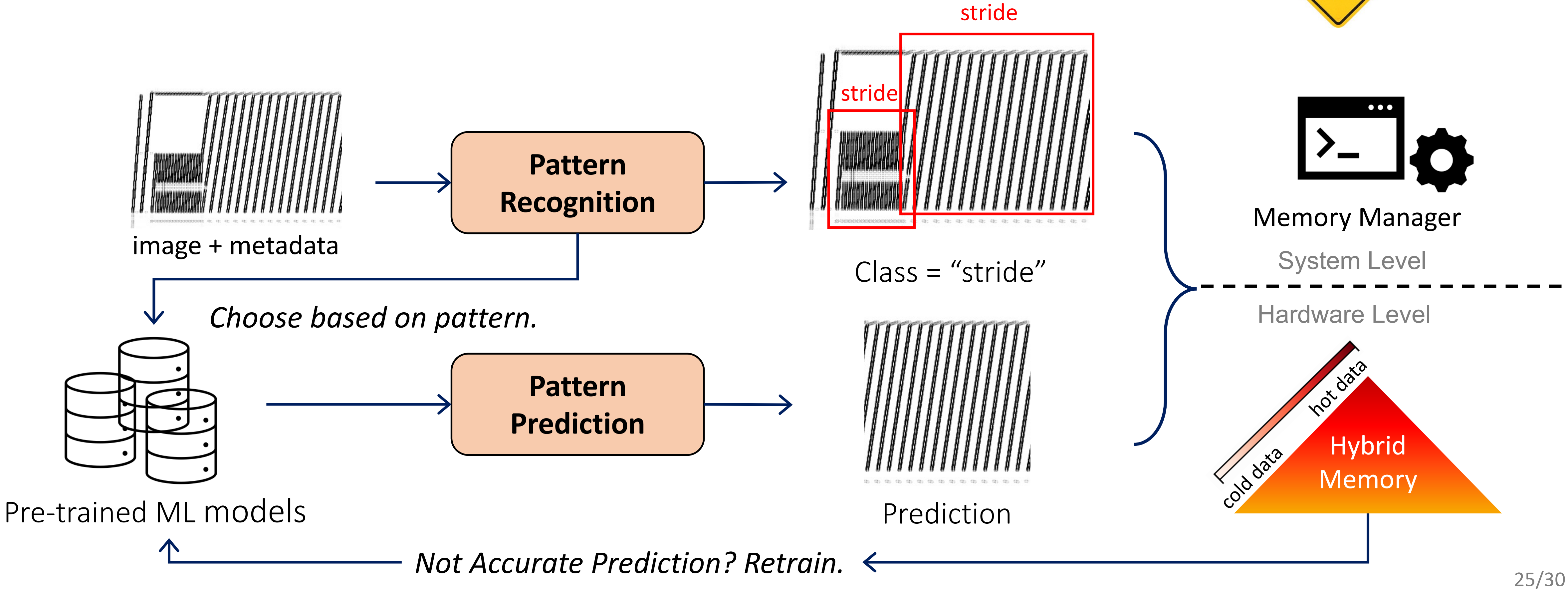
The image-based ConvLSTM makes more accurate predictions.

Computer Vision + Machine Learning for Systems (2)

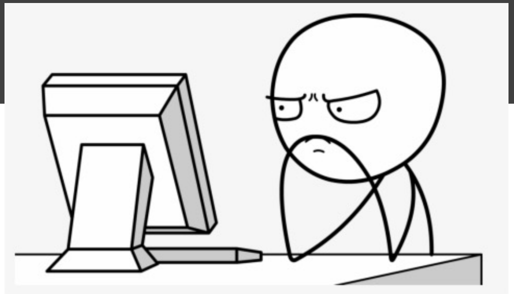
What can an image-based system pipeline look like?

WORK IN PROGRESS

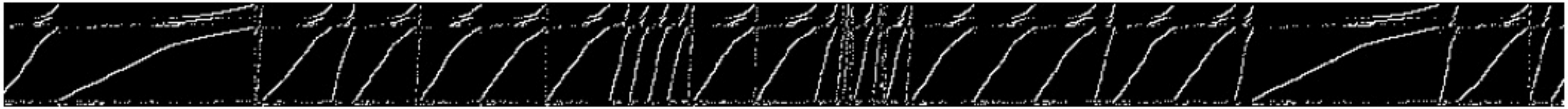
E.g., learning memory access patterns.



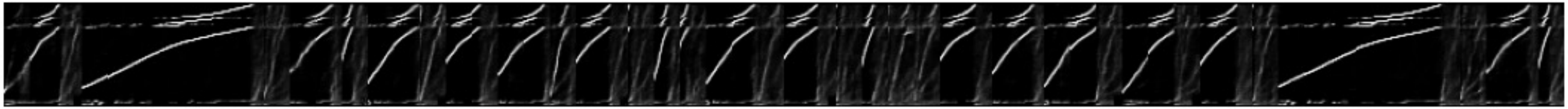
Early Results on Image-based Pattern Prediction



Ground Truth



Prediction



More challenging, since the data access patterns are more complex.

Talk Outline

Why do we need Smarter and Faster Systems?
The evolution of the hardware technologies, calls for software improvements.

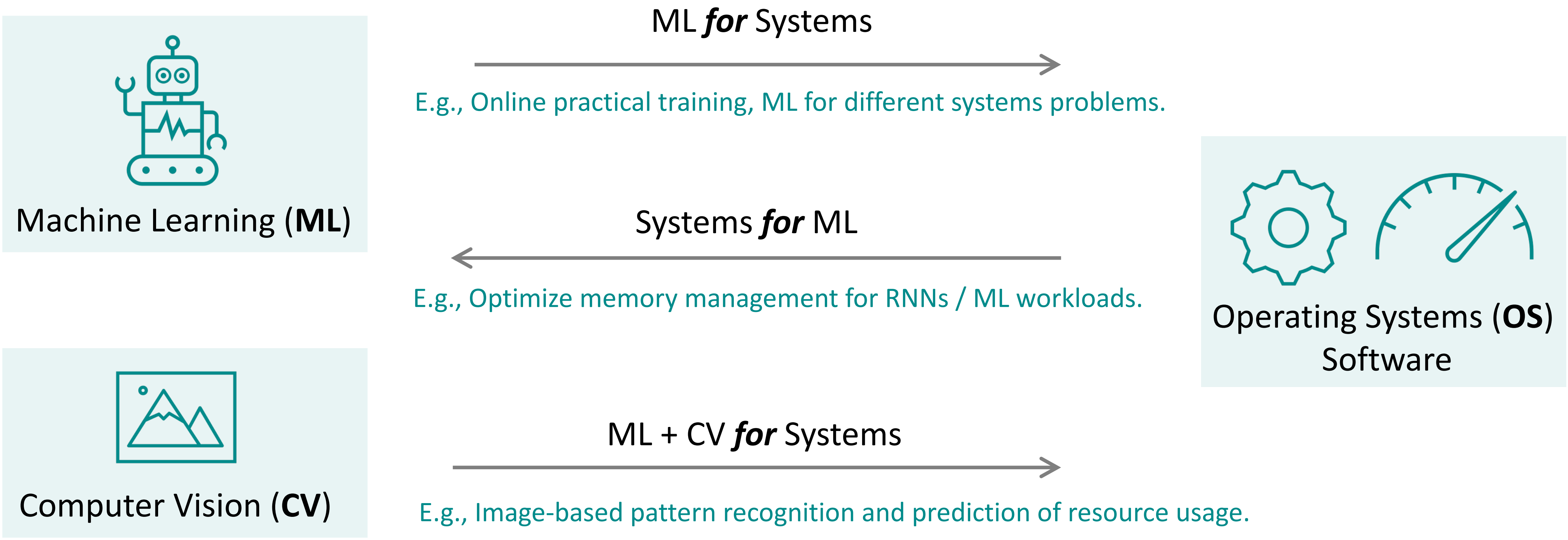
Building *Smart* Systems
Using machine and human intelligence to build practical ML-based systems.

Building *Fast* Systems
Reducing ML-based management overheads with visualization.
Building image-based system pipelines.

Future Research Directions

Future Research Directions

My research lies at the intersection of Machine Learning and Systems.



Machine Learning (**ML**)

Computer Vision (**CV**)

Operating Systems (**OS**)
Software

ML ***for*** Systems

E.g., Online practical training, ML for different systems problems.

Systems ***for*** ML

E.g., Optimize memory management for RNNs / ML workloads.

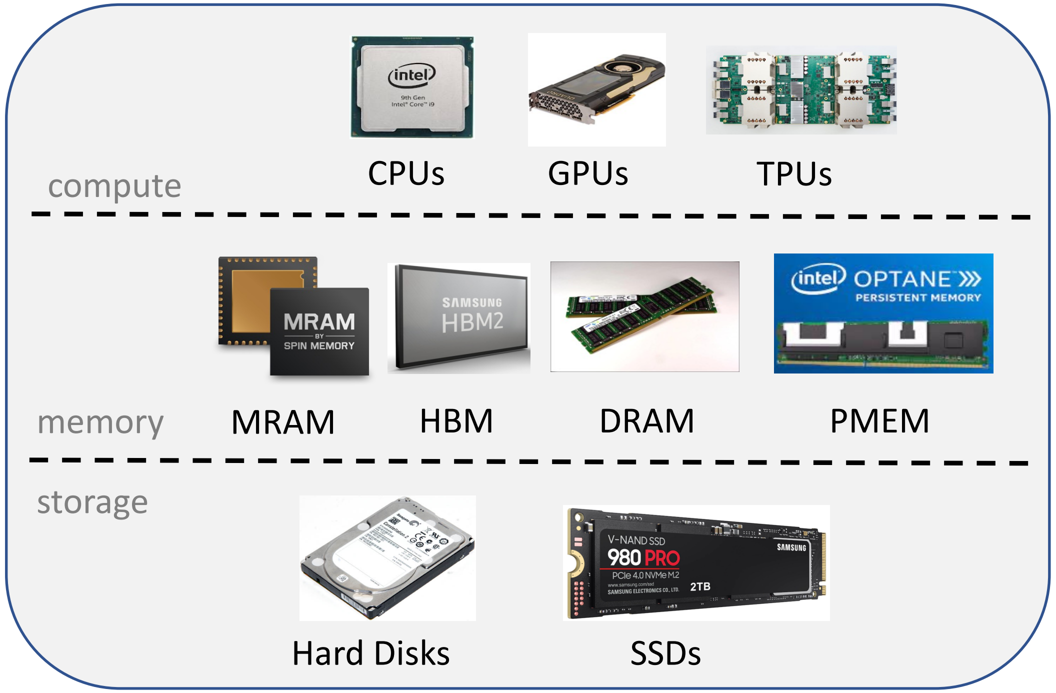
ML + CV ***for*** Systems

E.g., Image-based pattern recognition and prediction of resource usage.

Intelligent Management of Extreme Heterogeneity

Hardware configuration?

Data / Resource Management across layers / nodes?


compute

CPUs

GPUs

TPUs

memory

MRAM

HBM

DRAM

PMEM

storage

Hard Disks

SSDs

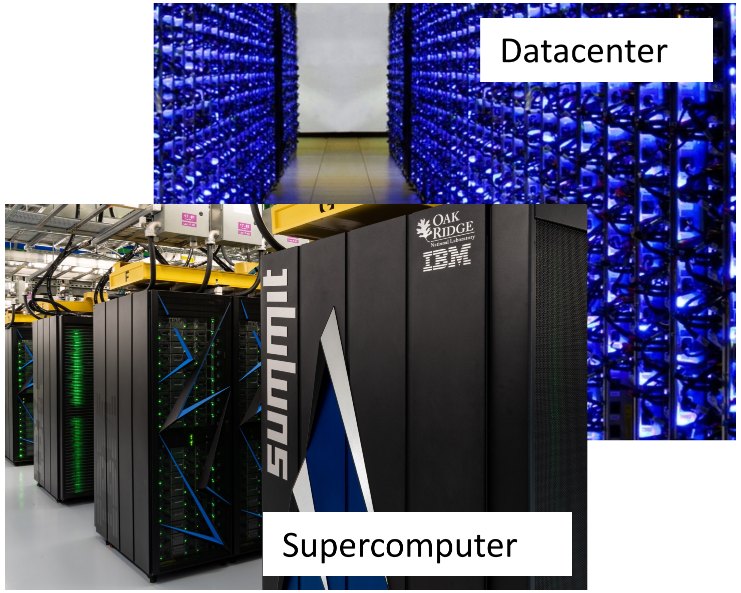
Local Node

Users

Multi-tenancy?
Isolation?

Performance?
Cost / Energy /
Resource Efficiency?

High-Speed Interconnects



Massive Node Clusters
Disaggregated Resources

System vs. HW / SW co-design?

ML integration Aspects:

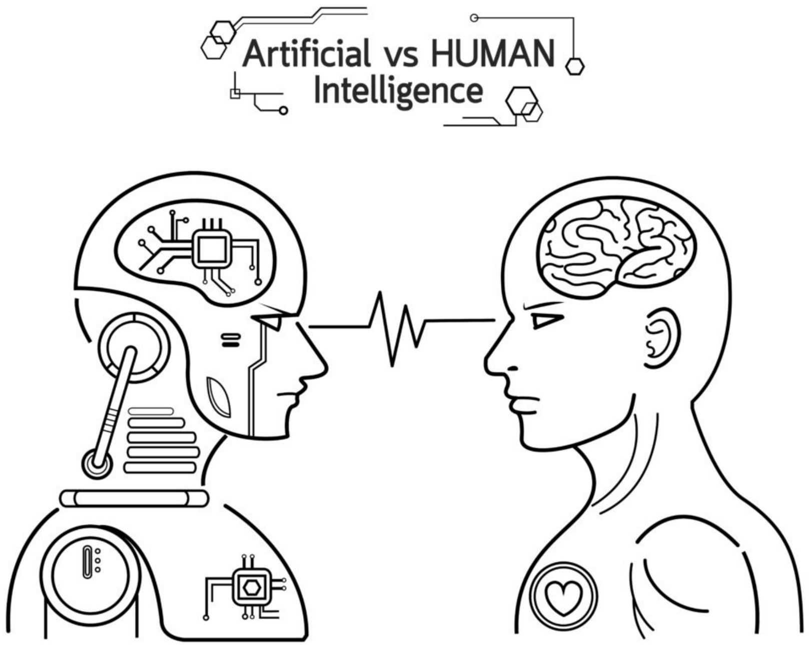
Necessity Effectiveness Practicality Interpretability

Scan this to find more about my work.

Smart

Fast

Systems



How can we use our human intelligence to build **practical** systems that leverage machine learning and computer vision?

Machine Learning

Computer Vision