Scalable Image AI via Self-Designing Storage

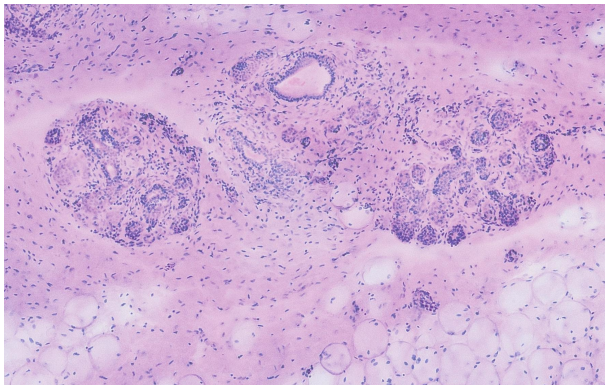
Utku Sirin

DASlab @ Harvard SEAS

work with Victoria Kauffman, Aadit Saluja, Florian Klein, Jeremy Hsu, Vlad Cainamisir, Qitong Wang, Konstantinos Kopsinis, Stratos Idreos

*What if we can **reason** about image data?*

Seeing is at center of AI



Digital Pathology

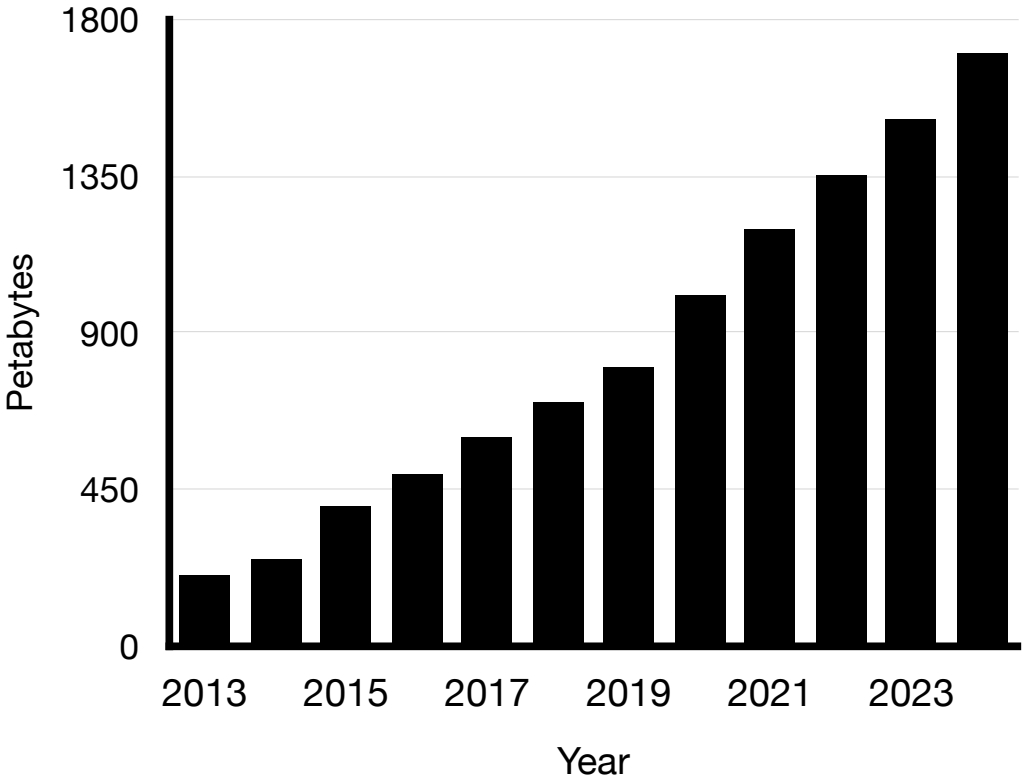
Textile Recycling

Robot Navigation

DASlab @ Harvard SEAS

Data size **Model size**

Data size

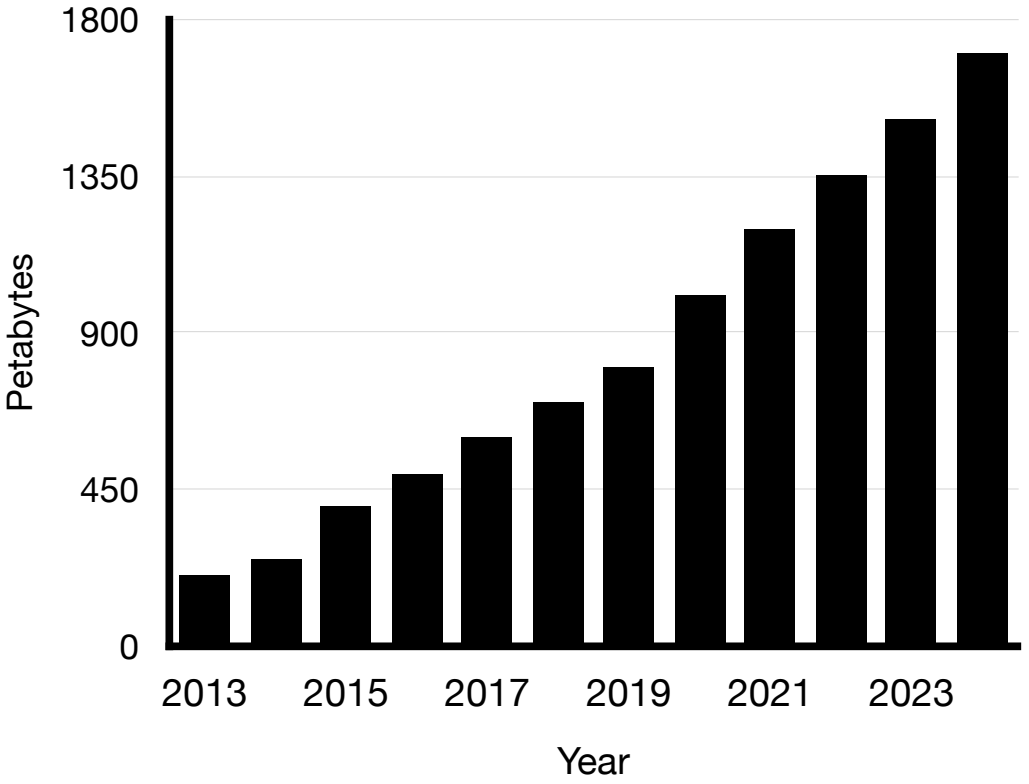


Model size

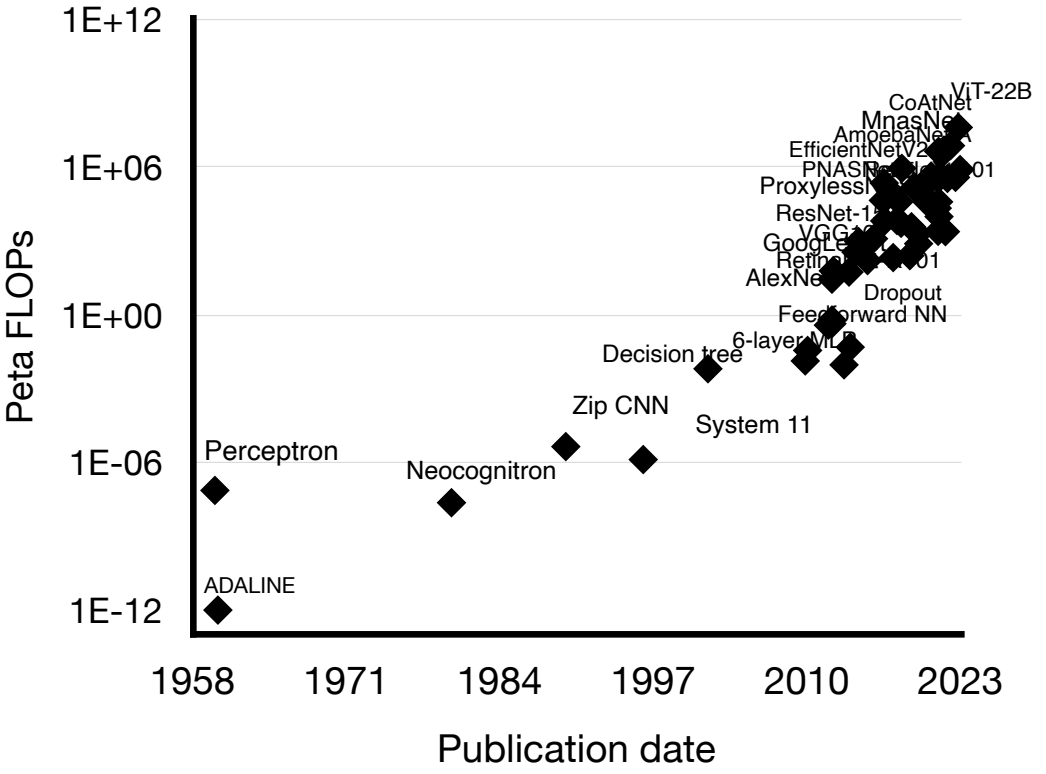
DASlab @ Harvard SEAS

[1] https://ourworldindata.org/artificial-intelligence
[2] https://ourworldindata.org/grapher/artificial-intelligence-training-computation

Data size

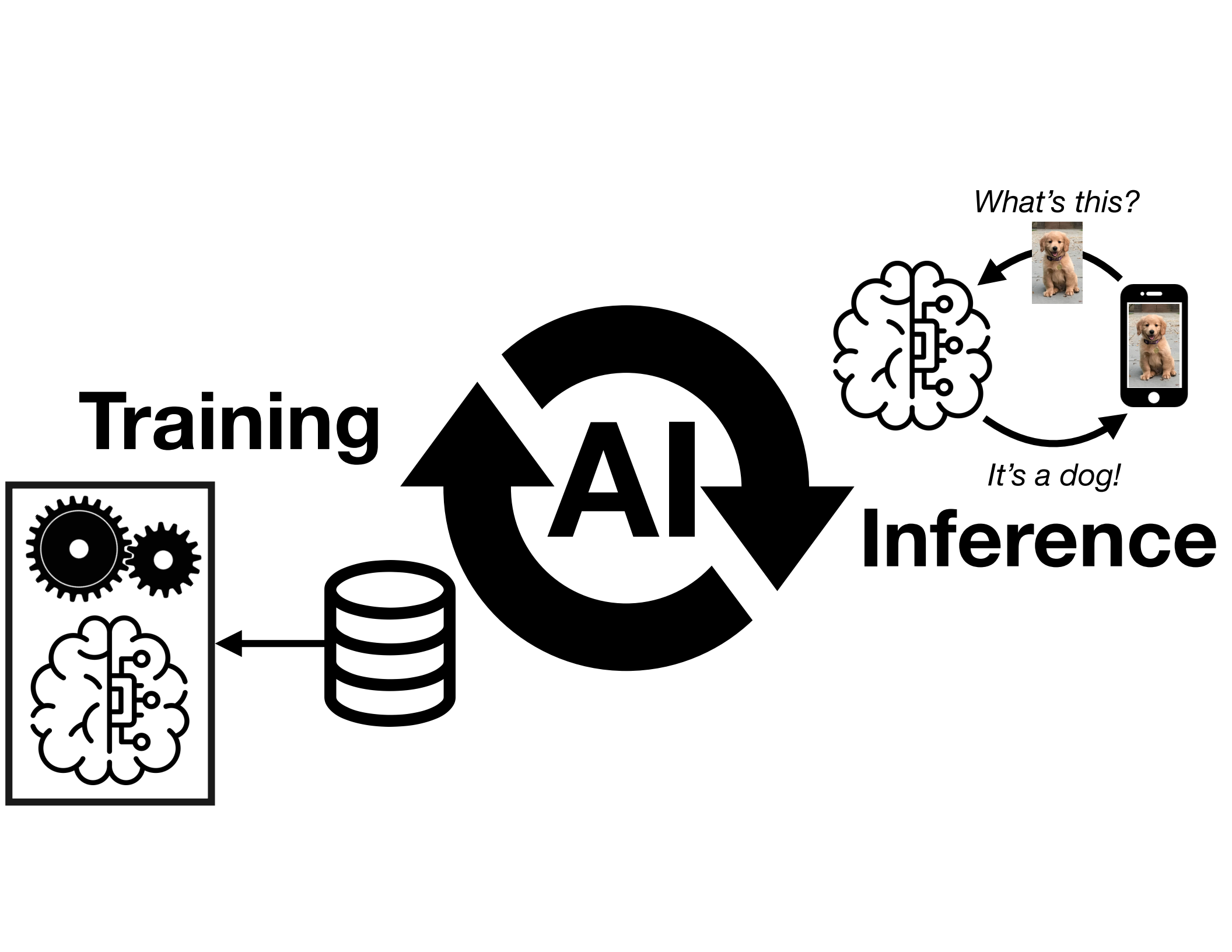


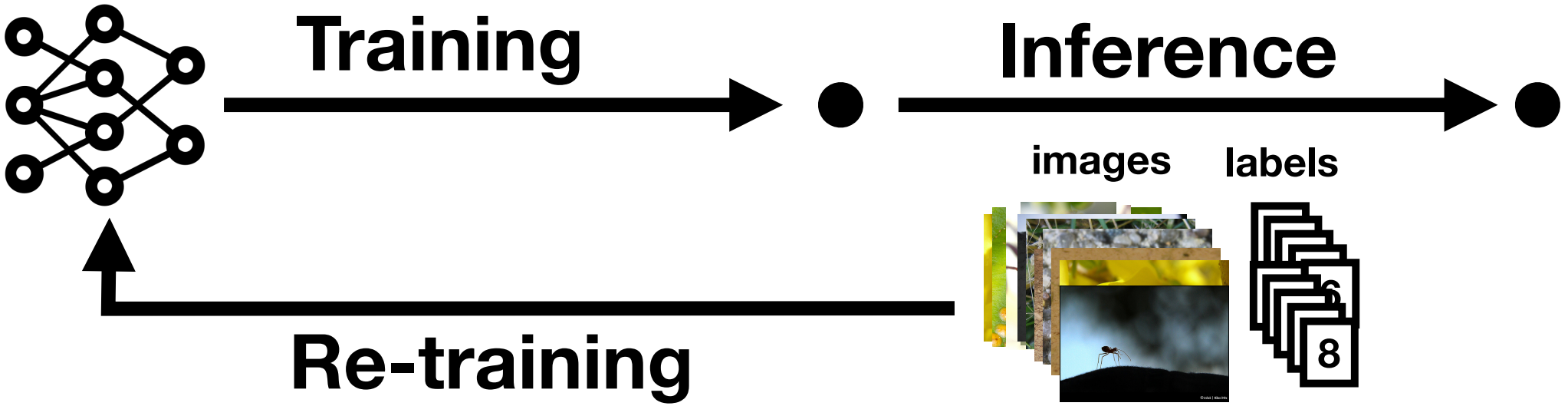
Model size



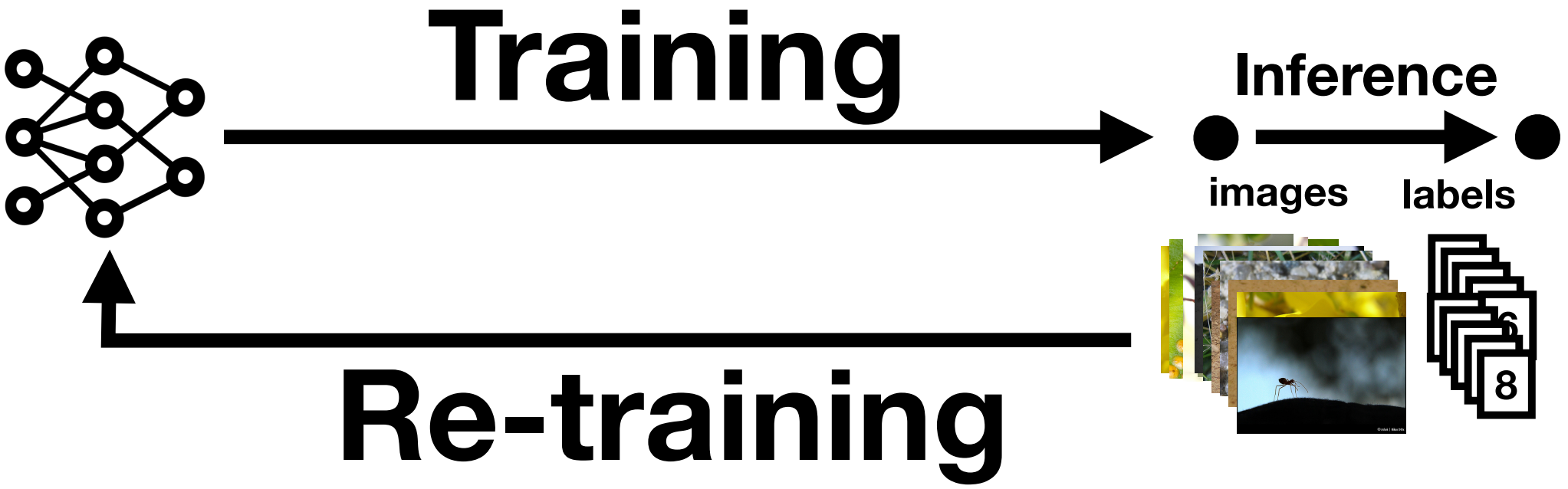
DASlab @ Harvard SEAS

[1] https://ourworldindata.org/artificial-intelligence
[2] https://ourworldindata.org/grapher/artificial-intelligence-training-computation

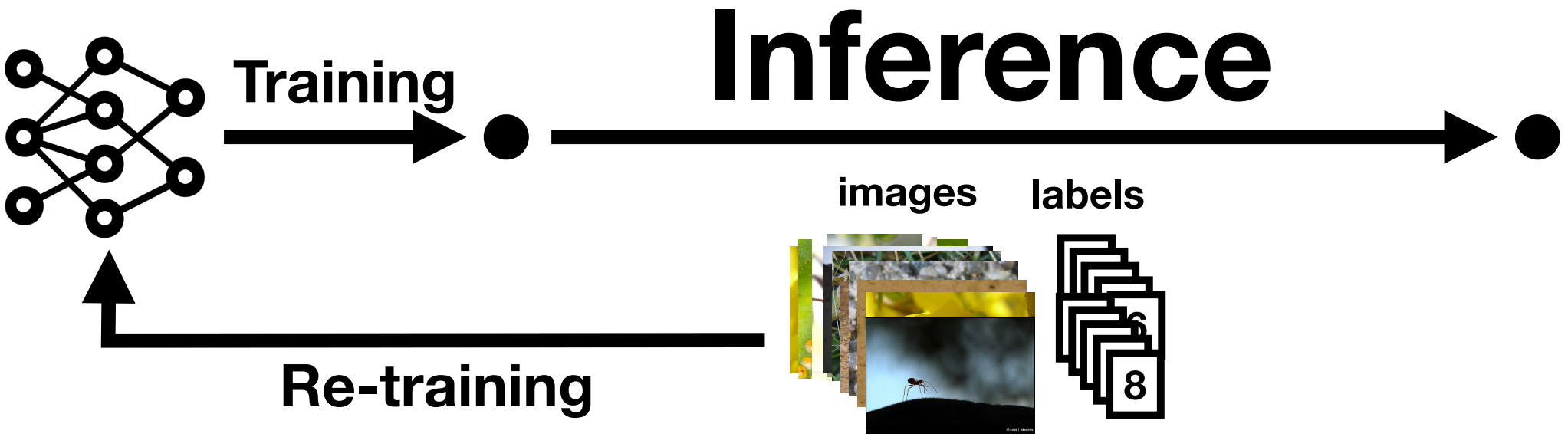

Training
AI
Inference
What's this?
It's a dog!



DASlab
@ Harvard SEAS

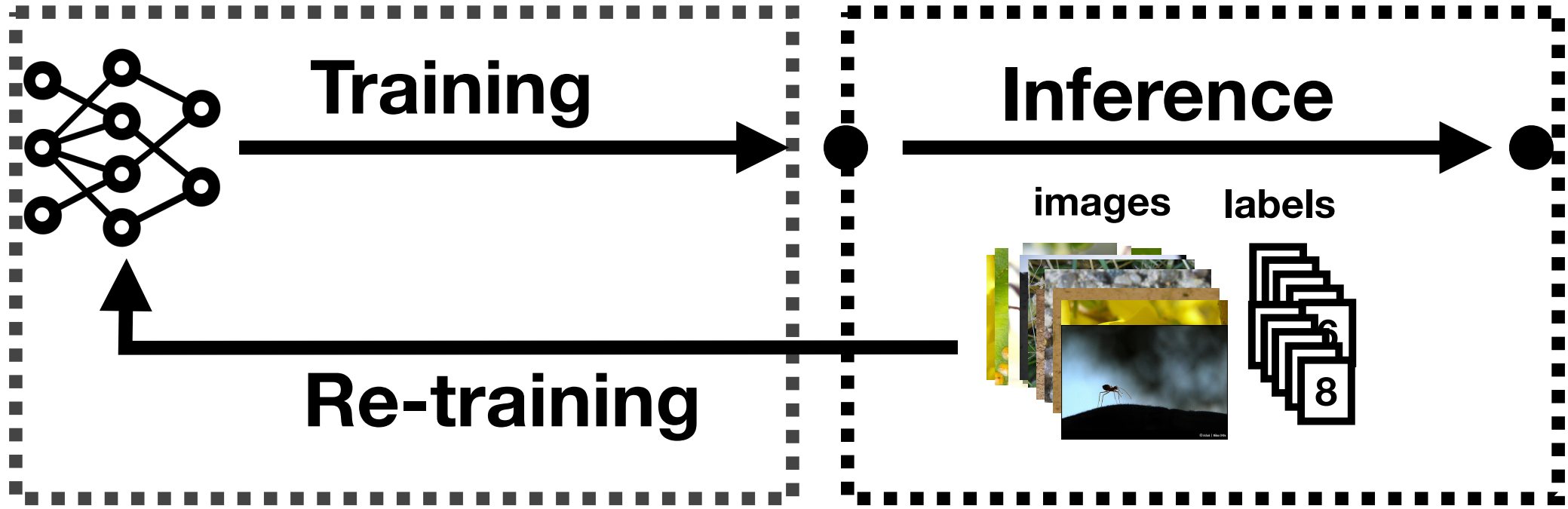


DASlab
@ Harvard SEAS



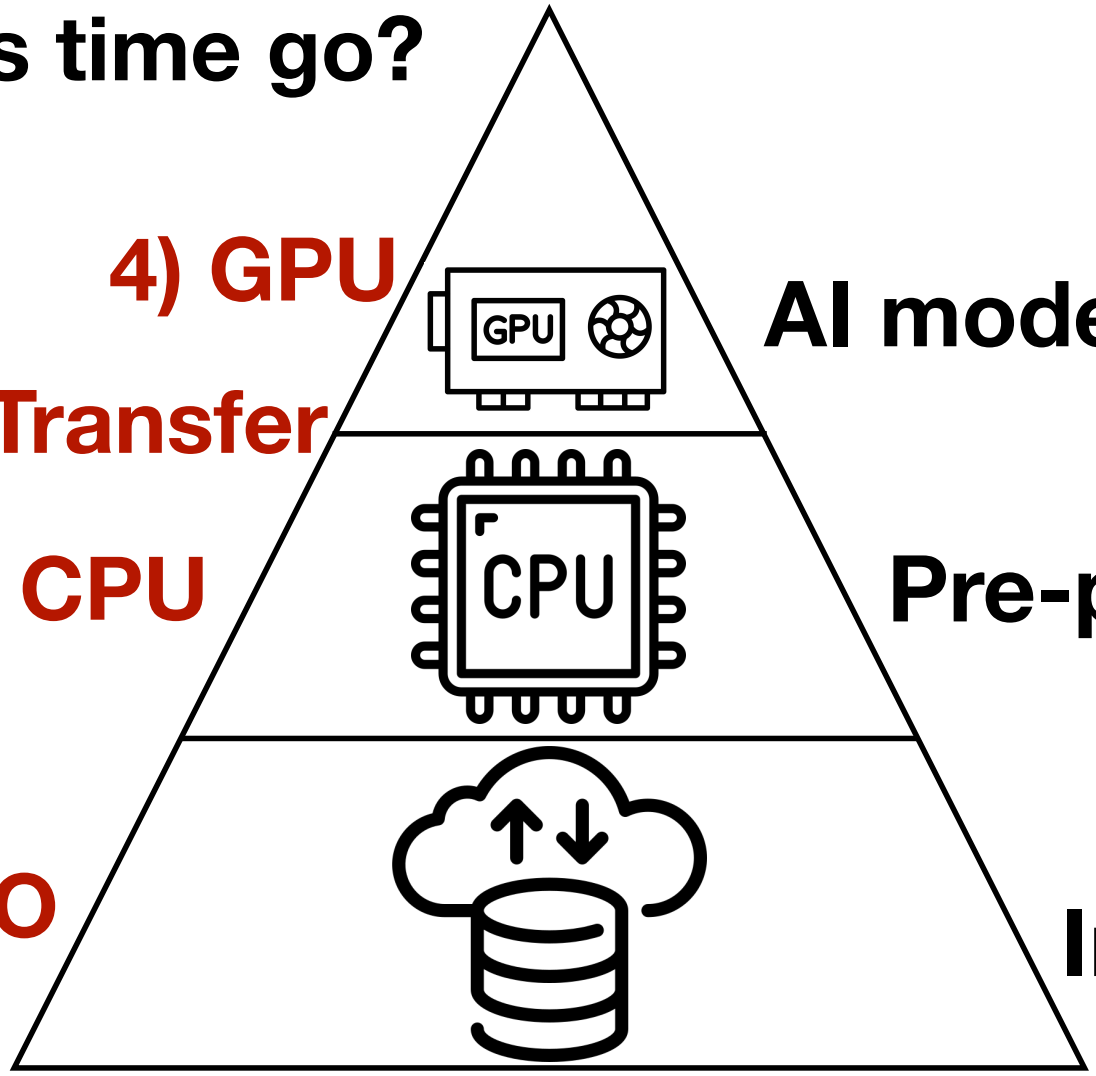
DASlab @ Harvard SEAS

[1] https://aws.amazon.com/blogs/aws/amazon-ec2-update-inf1-instances-with-aws-inferentia-chips-for-high-performance-cost-effective-inferencing/
[2] https://www.forbes.com/sites/moorinsights/2019/05/09/google-cloud-doubles-down-on-nvidia-gpus-for-inference/?sh=4c9817226792



DASlab
@ Harvard SEAS

Where does time go?



4) GPU

3) Transfer

2) CPU

1) Disk I/O

AI model

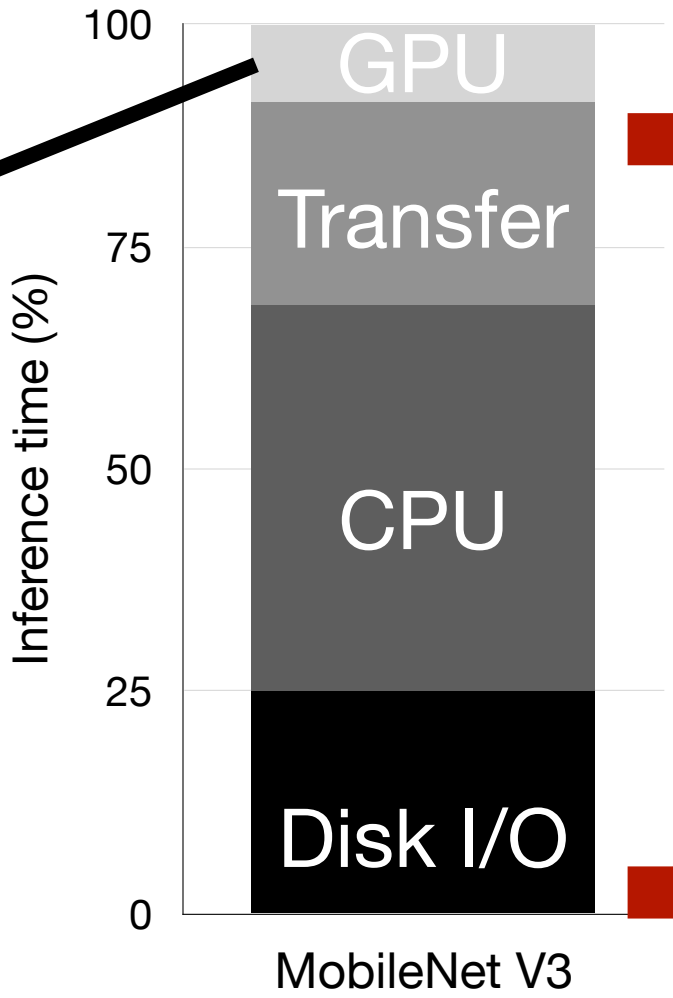
Pre-process

Image files

DASlab
@ Harvard SEAS

Where does time go?

Only 10% is GPU!

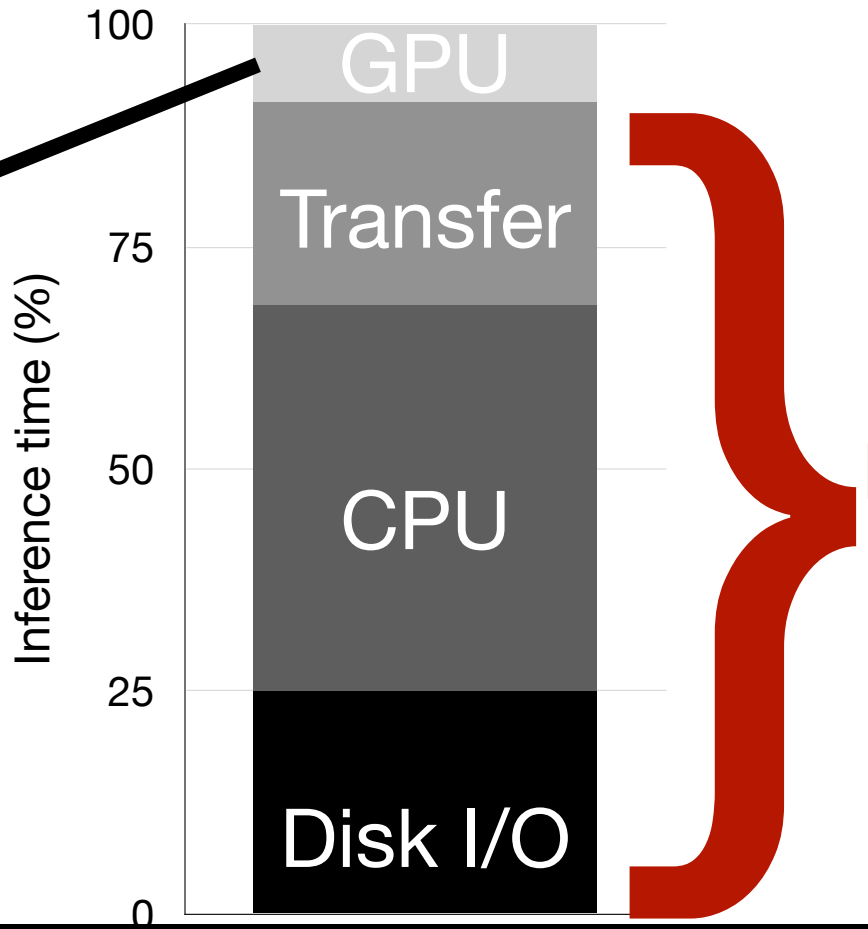


Data: ImageNet
AI Model: MobileNet V3
Machine: V100, PCIe Xeon, SSD
Framework: PyTorch v1

Data movement/ pre-processing

Where does time go?

Only 10% is GPU!



Data movement/ pre-processing

Data: ImageNet
AI Model: MobileNet V3
Machine: V100, PCIe Xeon, SSD
Framework: PyTorch v1

Re-consider Storage for AI

How do machines store images today?

JPEG

Joint Photographic Experts Group

standard

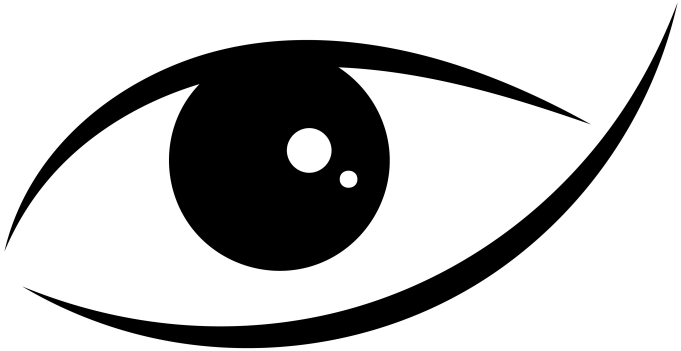
compression

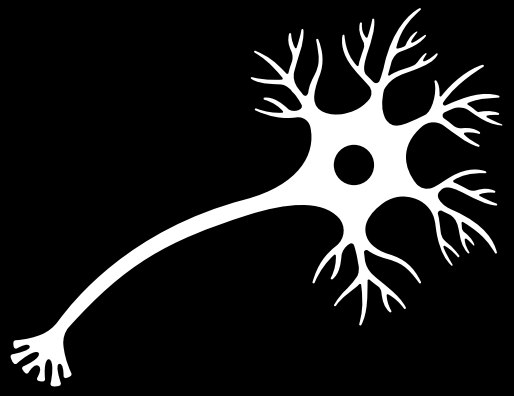
DASlab
@ Harvard SEAS

RAW

JPEG

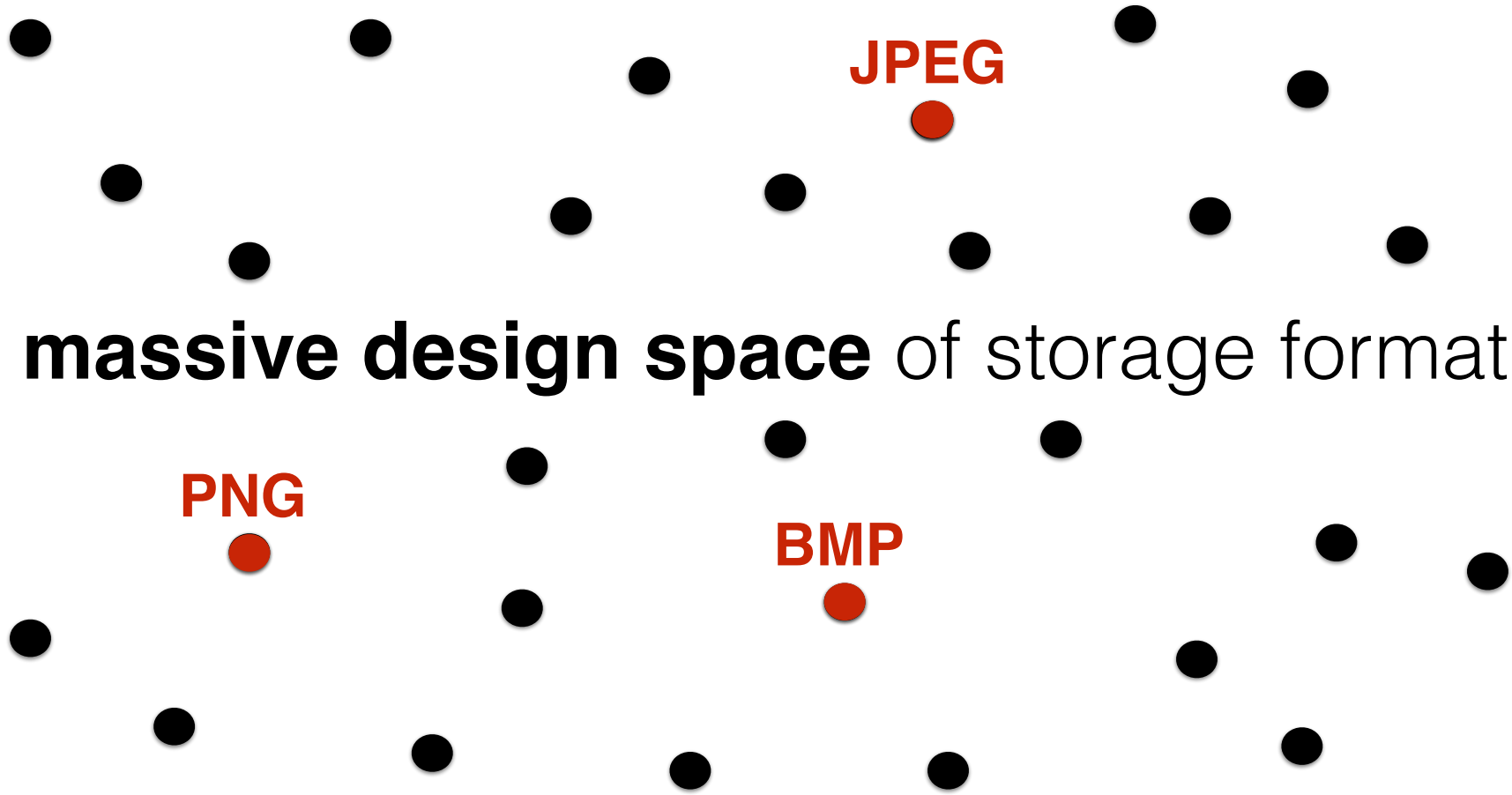
JPEG is designed for human eye





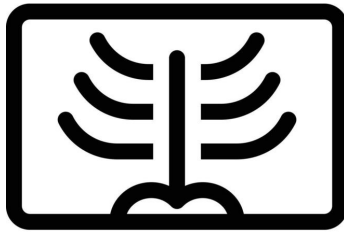
AI algorithms “see” images

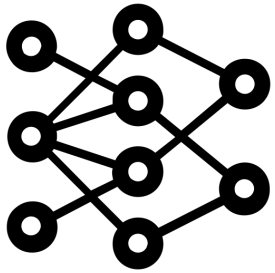
A few existing designs



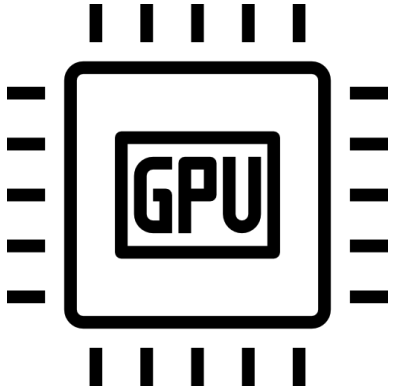
massive design space of storage formats

DASlab @ Harvard SEAS





massive space of AI problems



DASlab
@ Harvard SEAS

Thesis

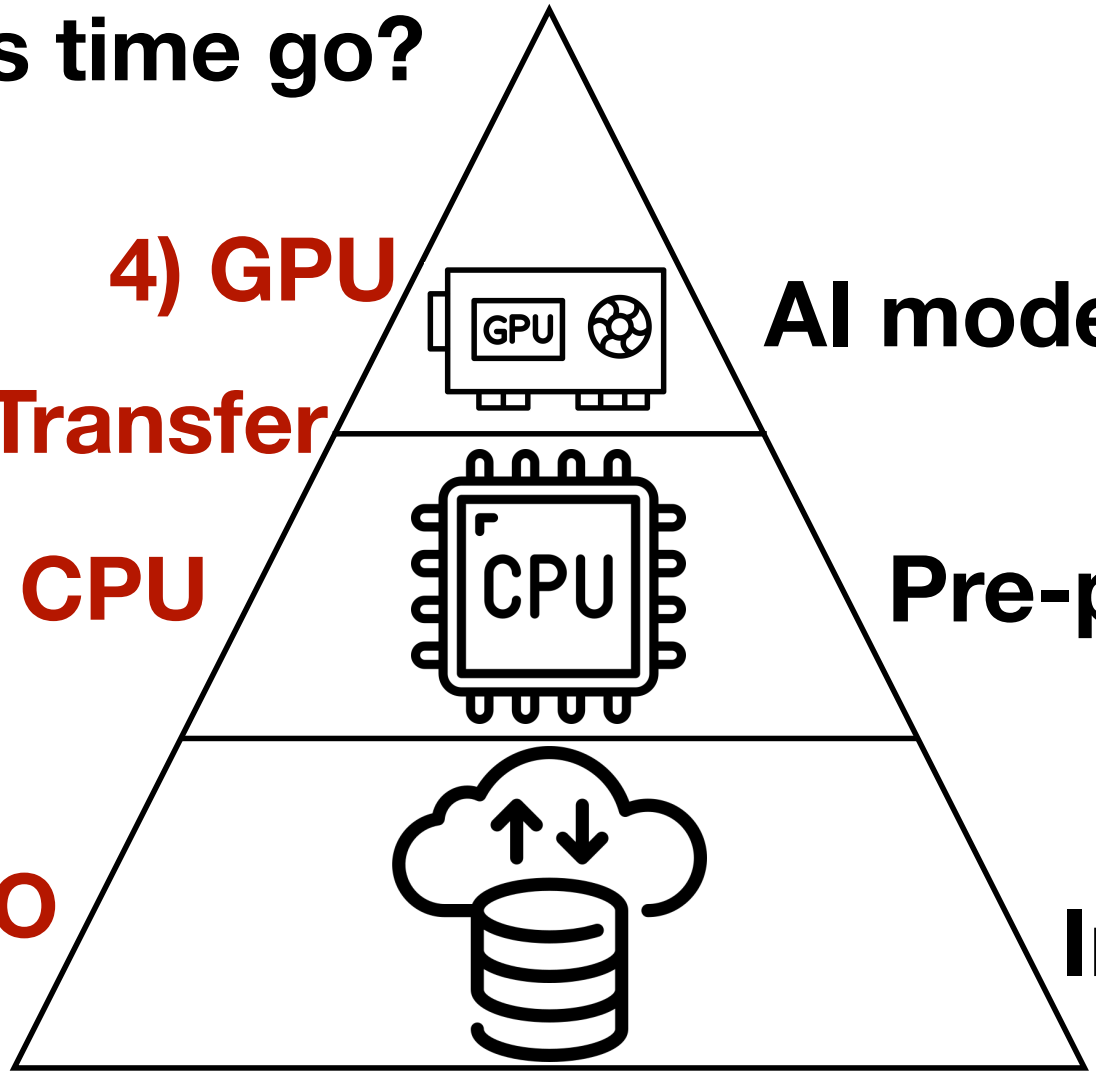
Storage determines end-to-end cost. Using a single storage for all problems results in inefficient AI systems. Orders of magnitude higher performance is only possible if the storage is tailored to the AI problem.

Image Calculator

A self-designing storage format that shapes itself for a given AI problem

Recap from previous class

Where does time go?



4) GPU

3) Transfer

2) CPU

1) Disk I/O

AI model

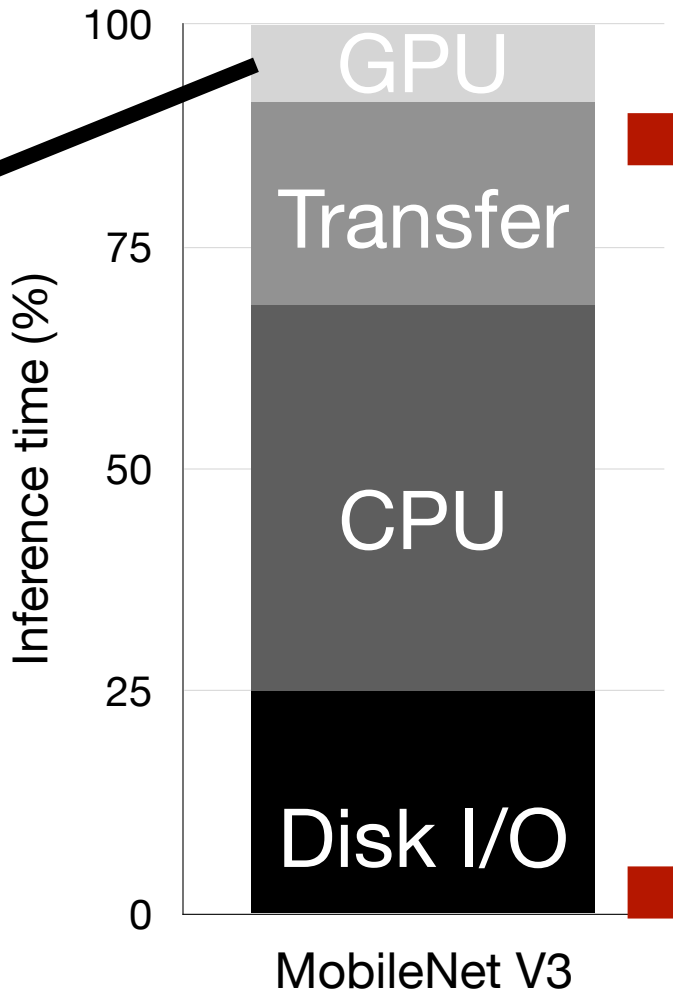
Pre-process

Image files

DASlab
@ Harvard SEAS

Where does time go?

Only 10% is GPU!

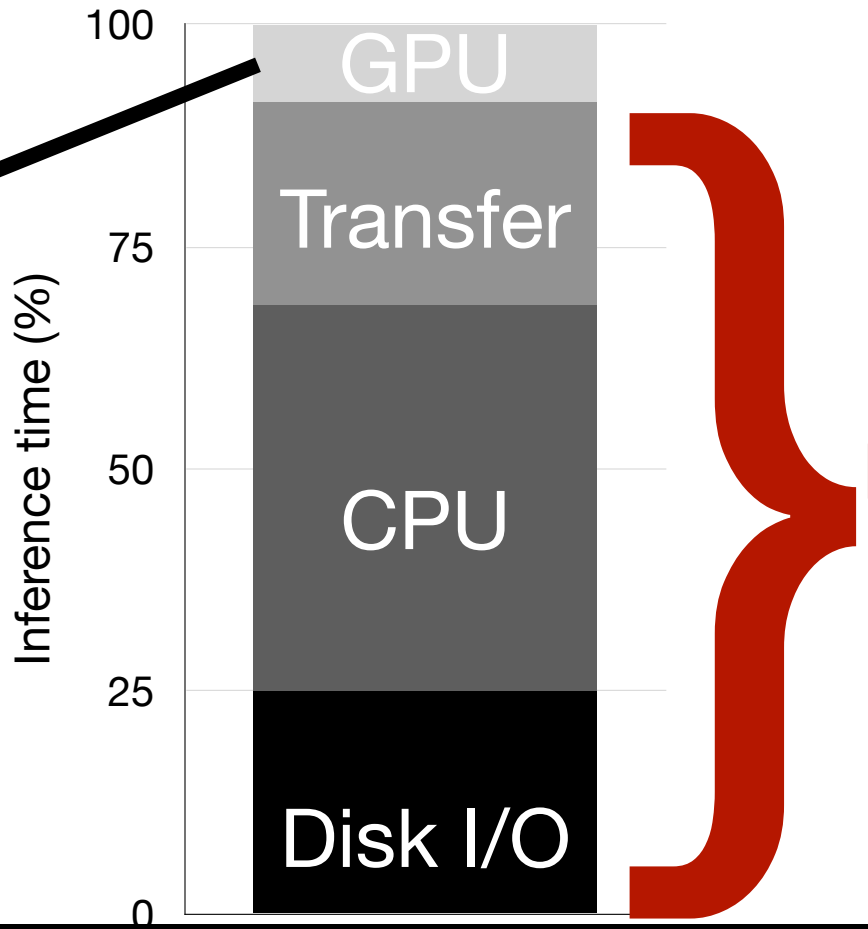


Data: ImageNet
AI Model: MobileNet V3
Machine: V100, PCIe Xeon, SSD
Framework: PyTorch v1

Data movement/ pre-processing

Where does time go?

Only 10% is GPU!

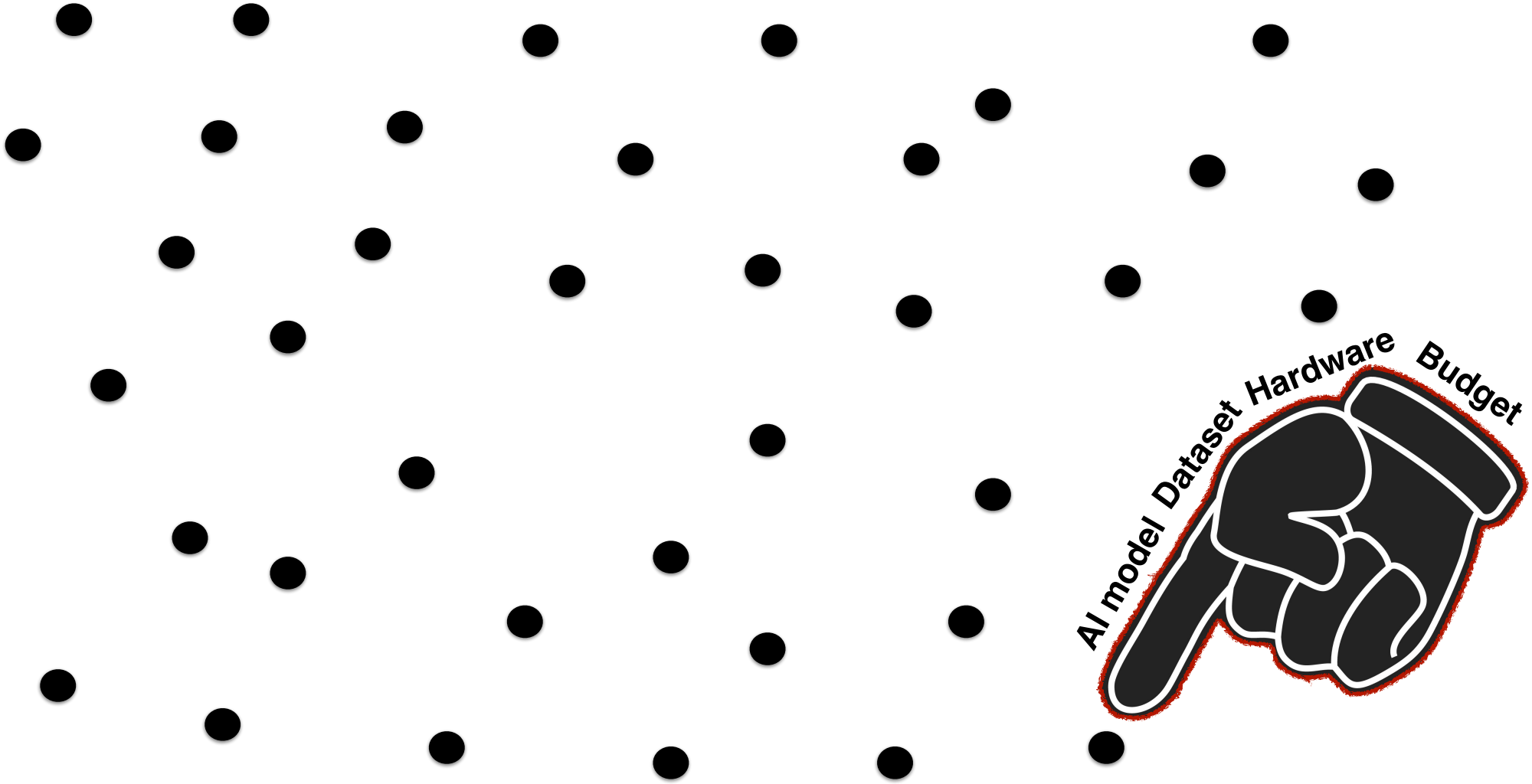


Data movement/ pre-processing

Data: ImageNet
AI Model: MobileNet V3
Machine: V100, PCIe Xeon, SSD
Framework: PyTorch v1

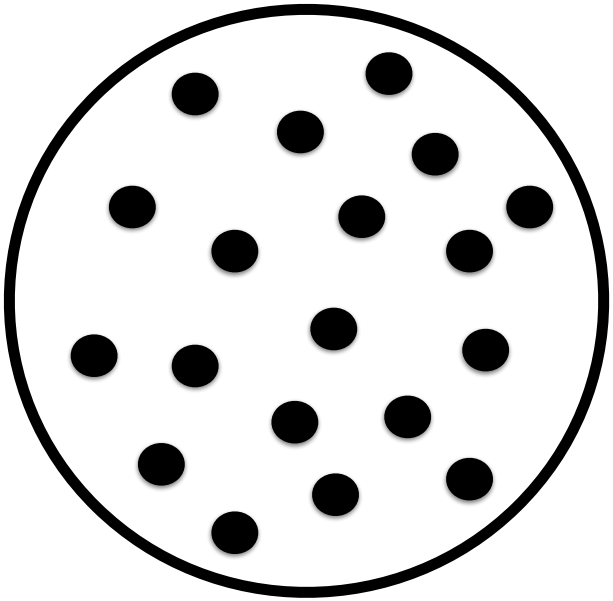
Re-consider Storage for AI

A self-designing File format

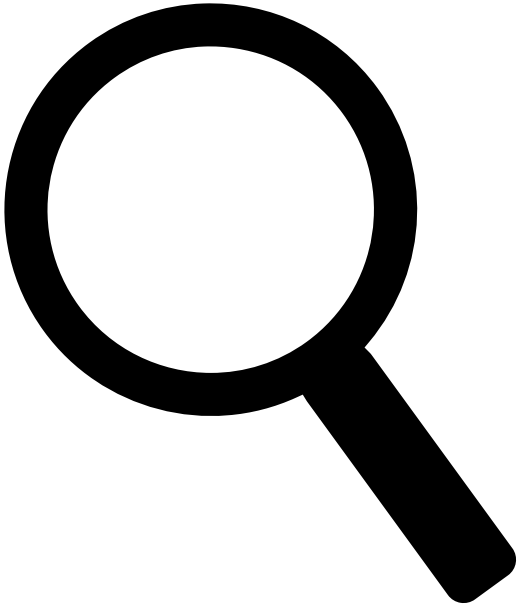


DASlab
@ Harvard SEAS

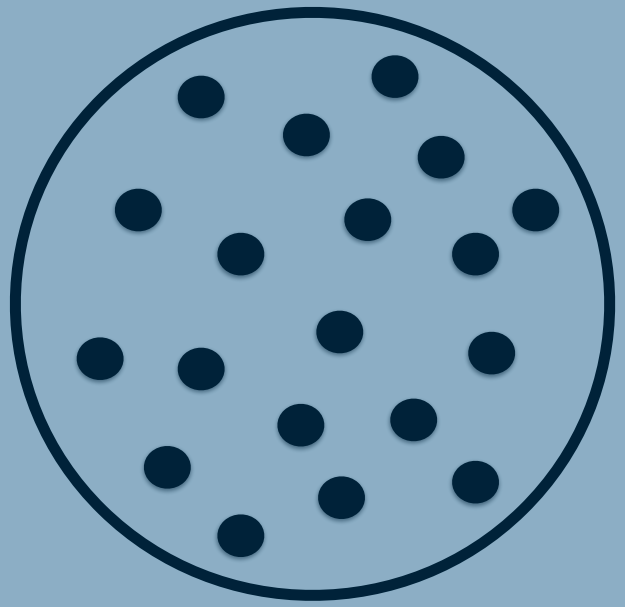
Design space



Search

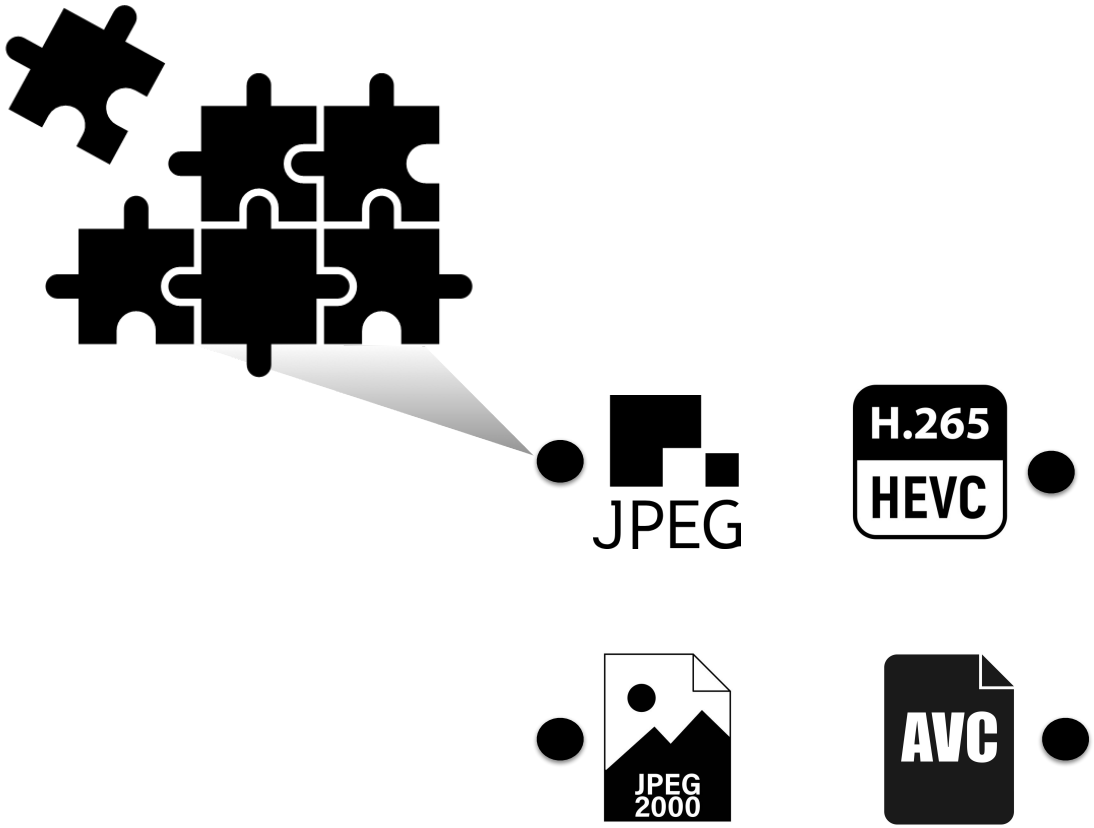


Design space

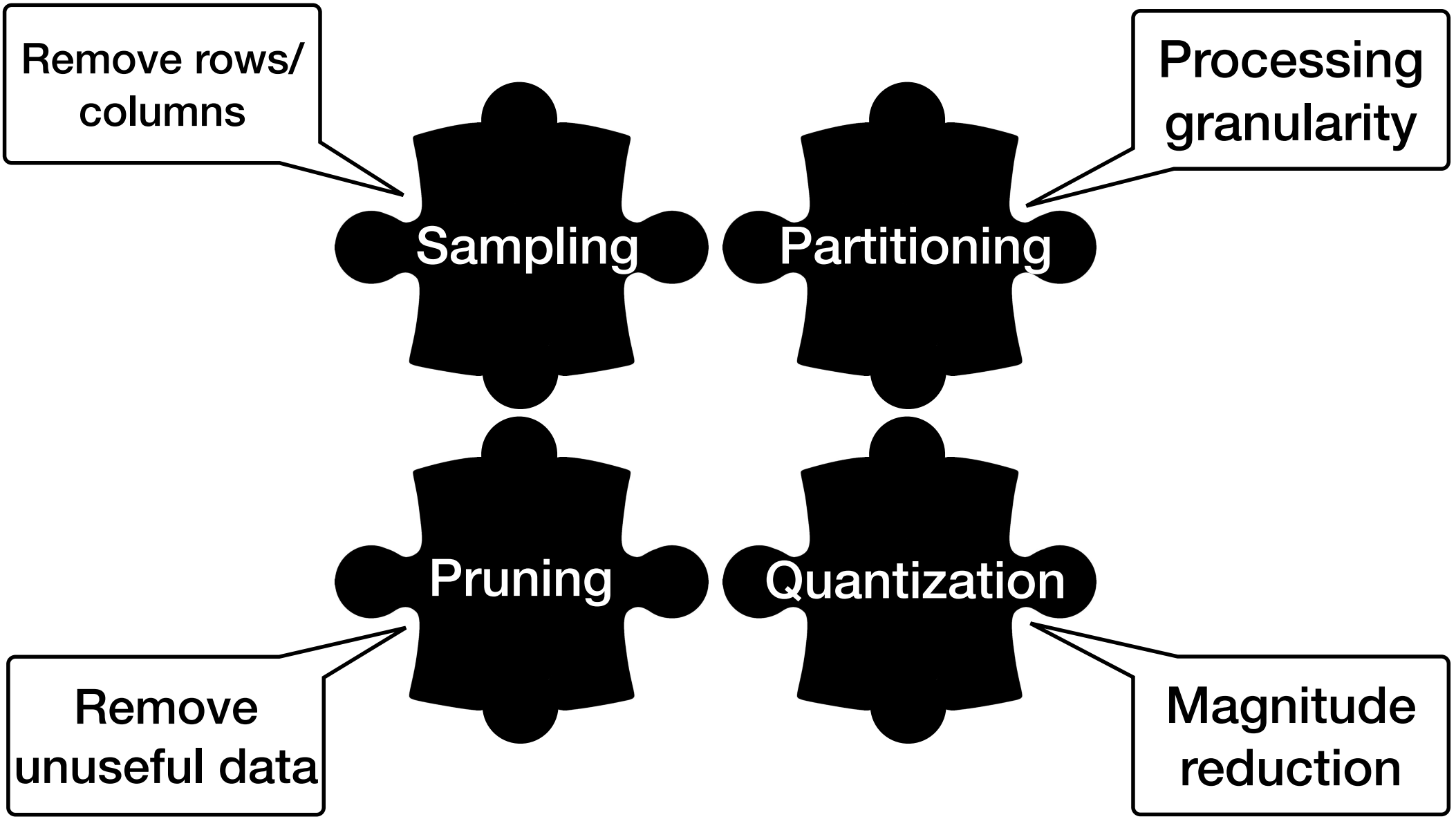


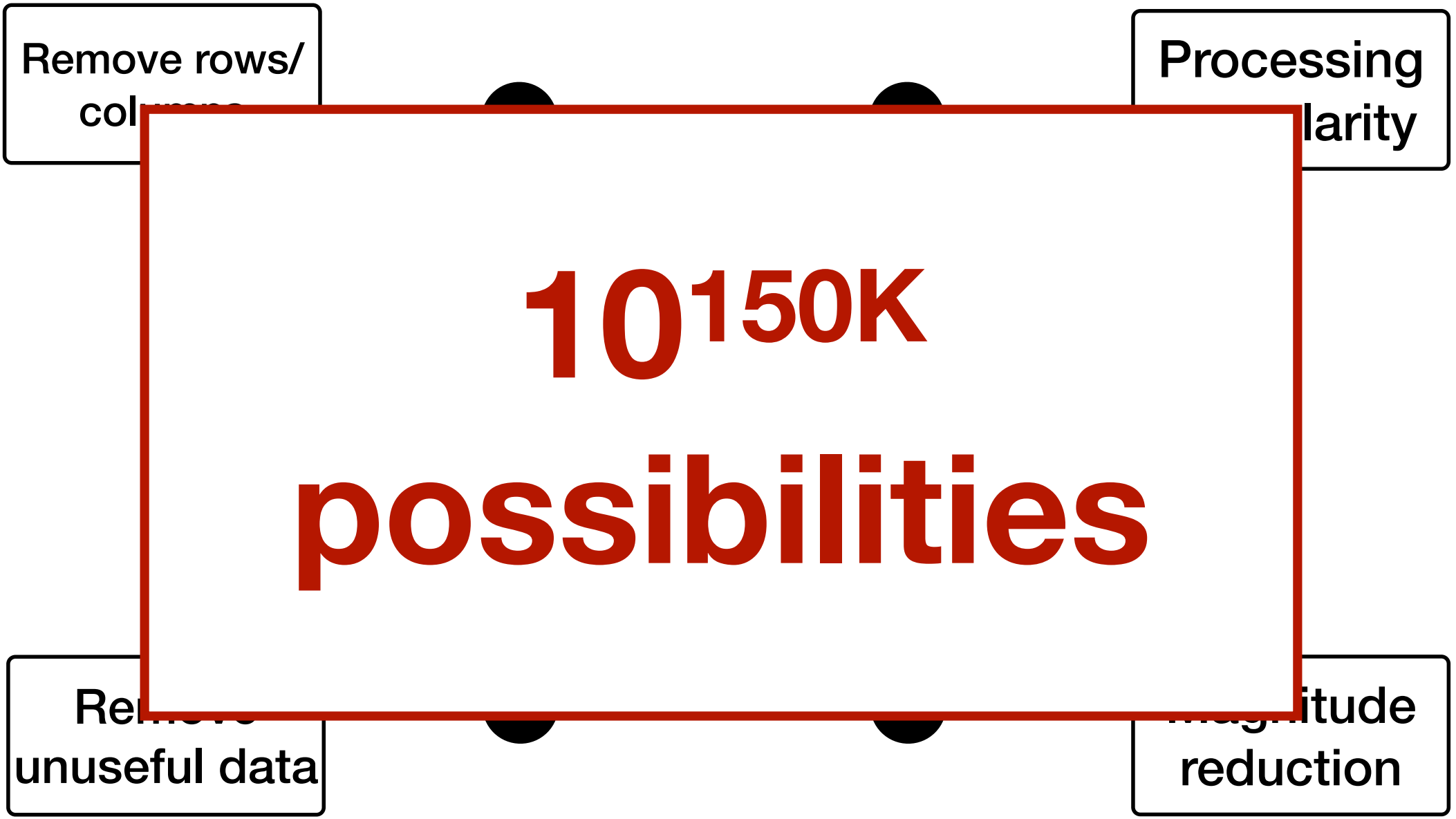
Search

DASlab
@ Harvard SEAS



DASlab
@ Harvard SEAS

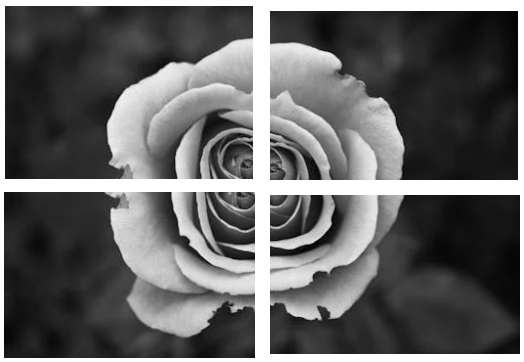

Remove rows/ columns
Sampling
Partitioning
Processing granularity
Pruning
Quantization
Remove unuseful data
Magnitude reduction


Remove rows/
col
Processing
larity
10^{150K}
possibilities
Re
unuseful data
itude
reduction

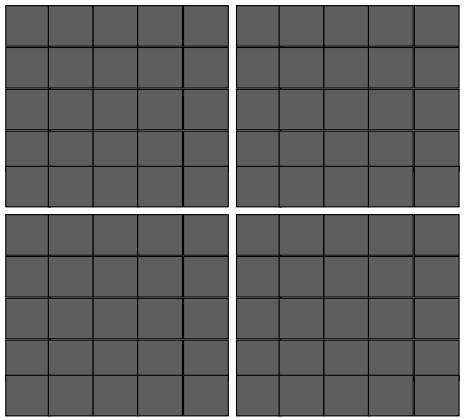
Designing A Domain: Pruning

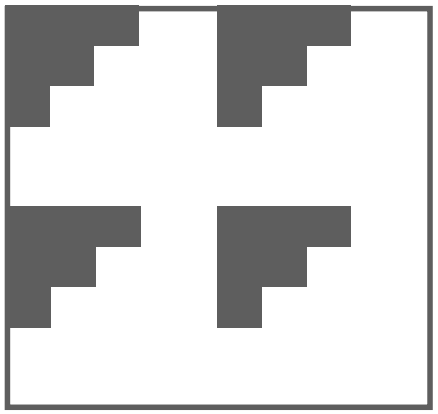
Background: Image Storage

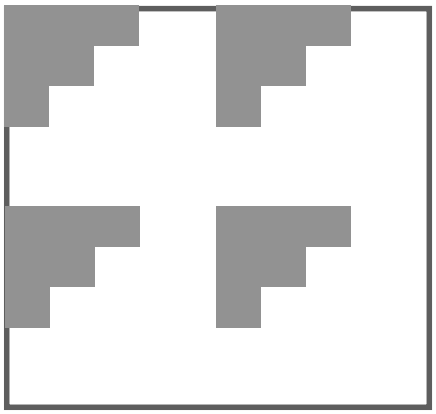
Partition → Transform → Prune → Quantize



Frequency Domain

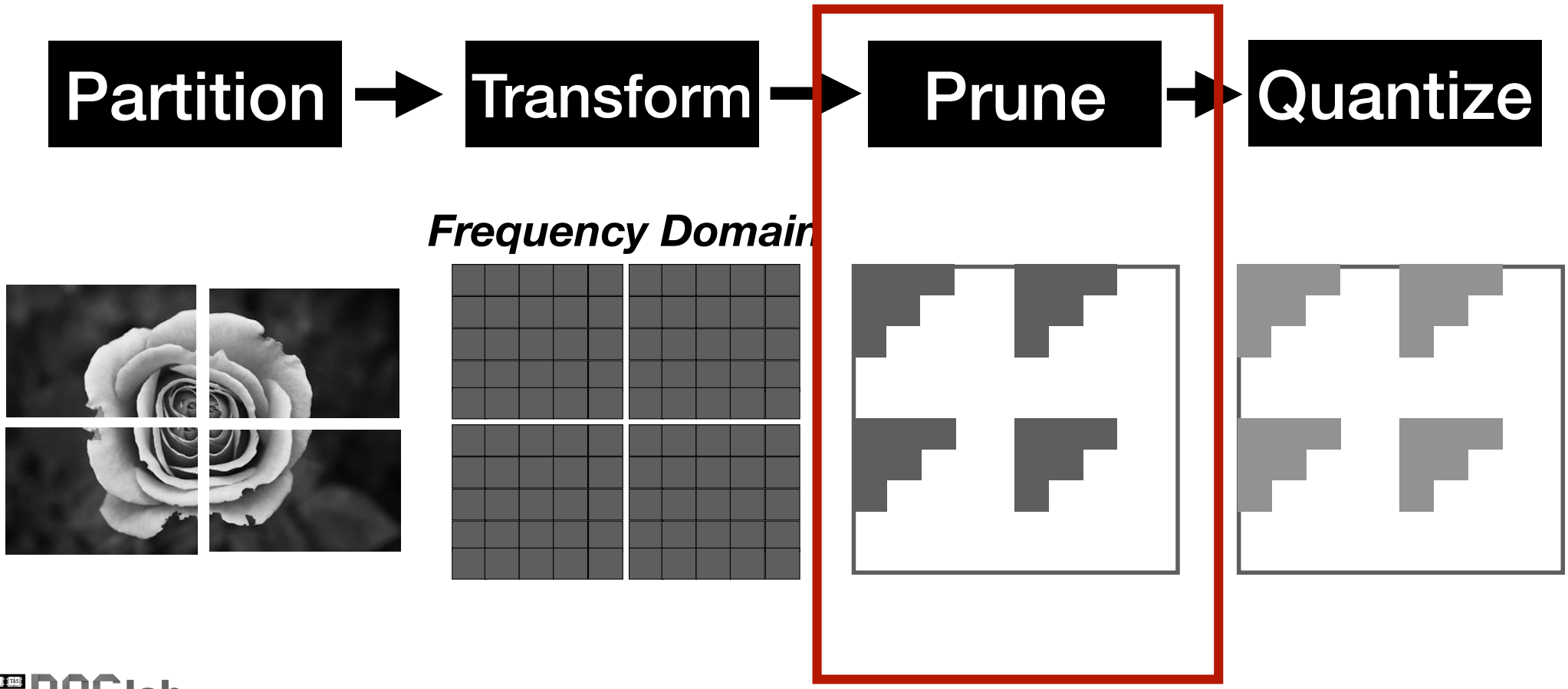


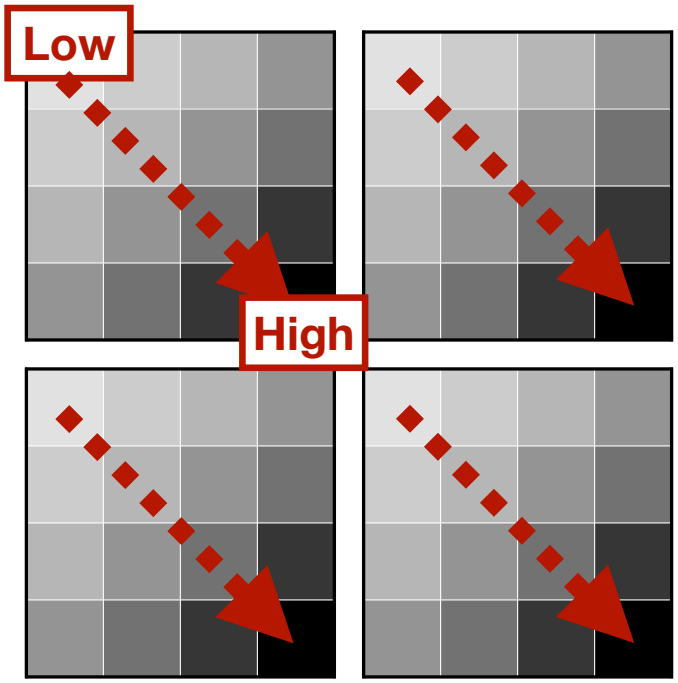


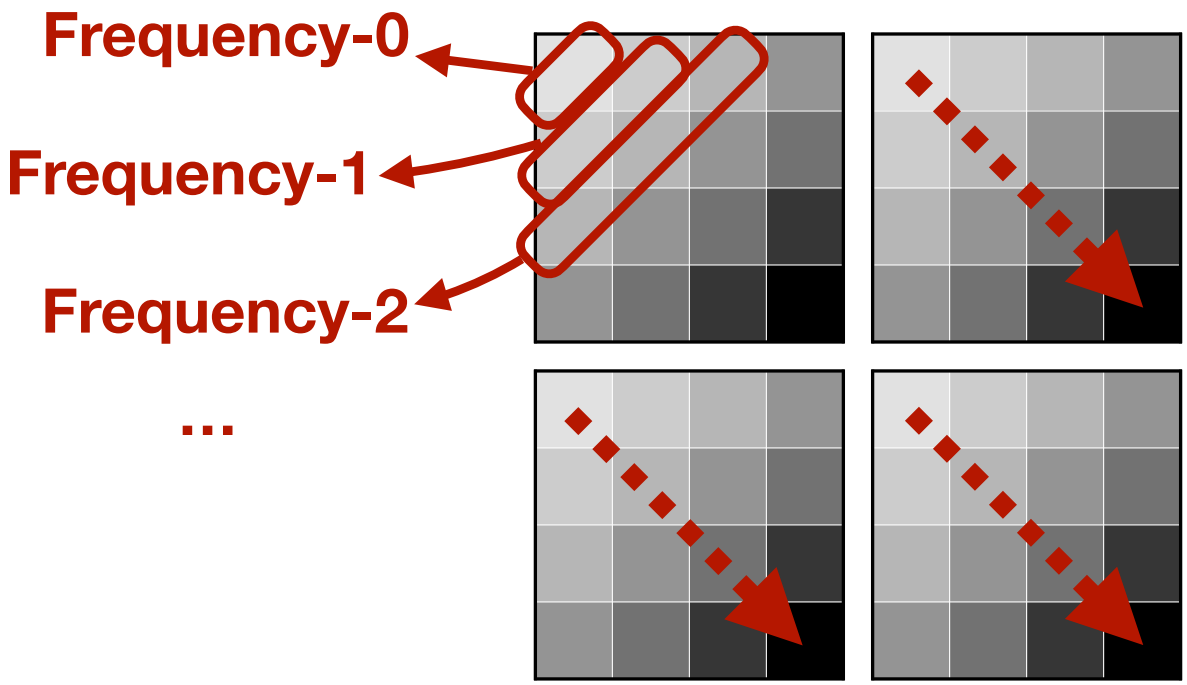


DASlab
@ Harvard SEAS

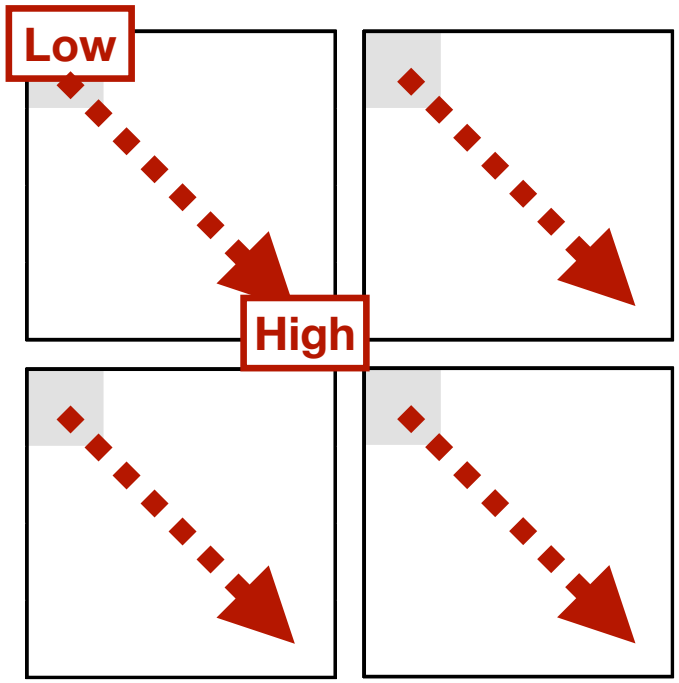
Background: Image Storage



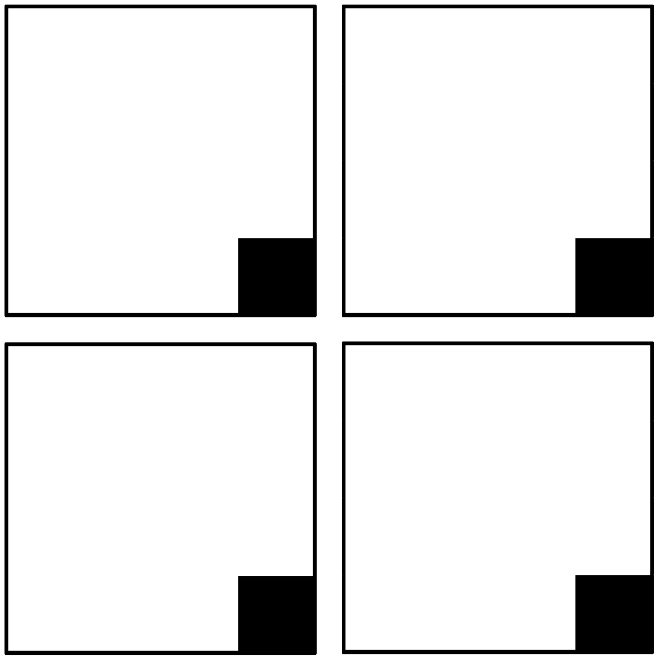

Low
High


Frequency-0
Frequency-1
Frequency-2
...

(1) High to low

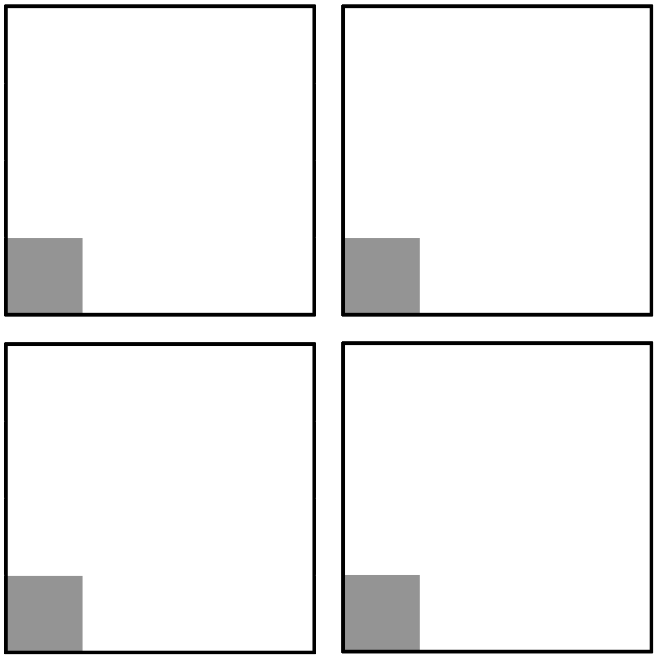


(2) Low to high

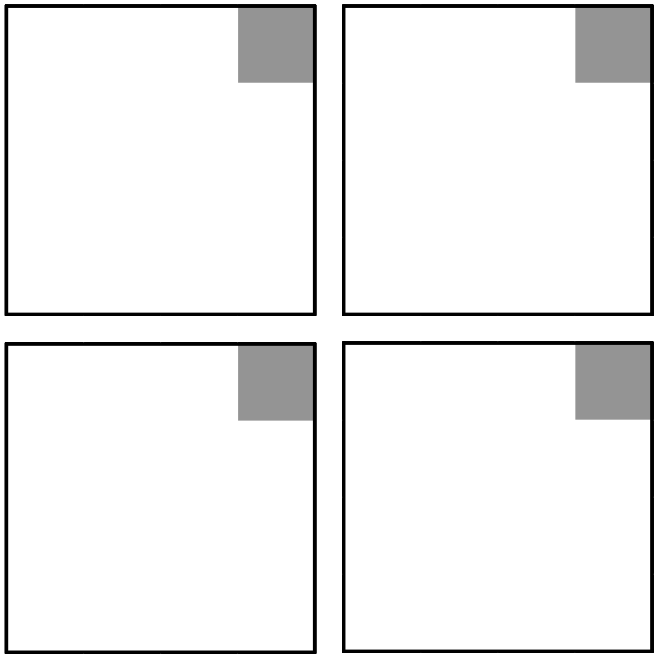


DASlab
@ Harvard SEAS

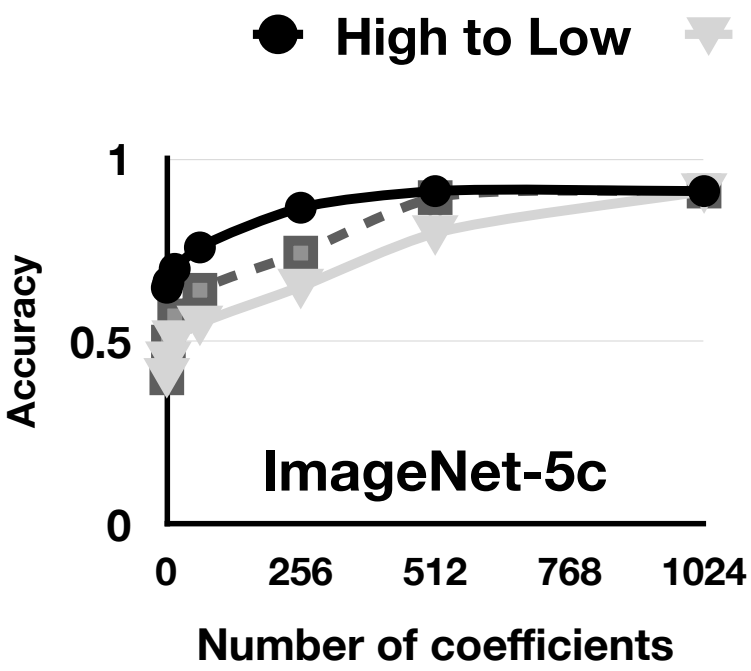
(3) Low & high — Horizontal



(4) Low & high — Vertical

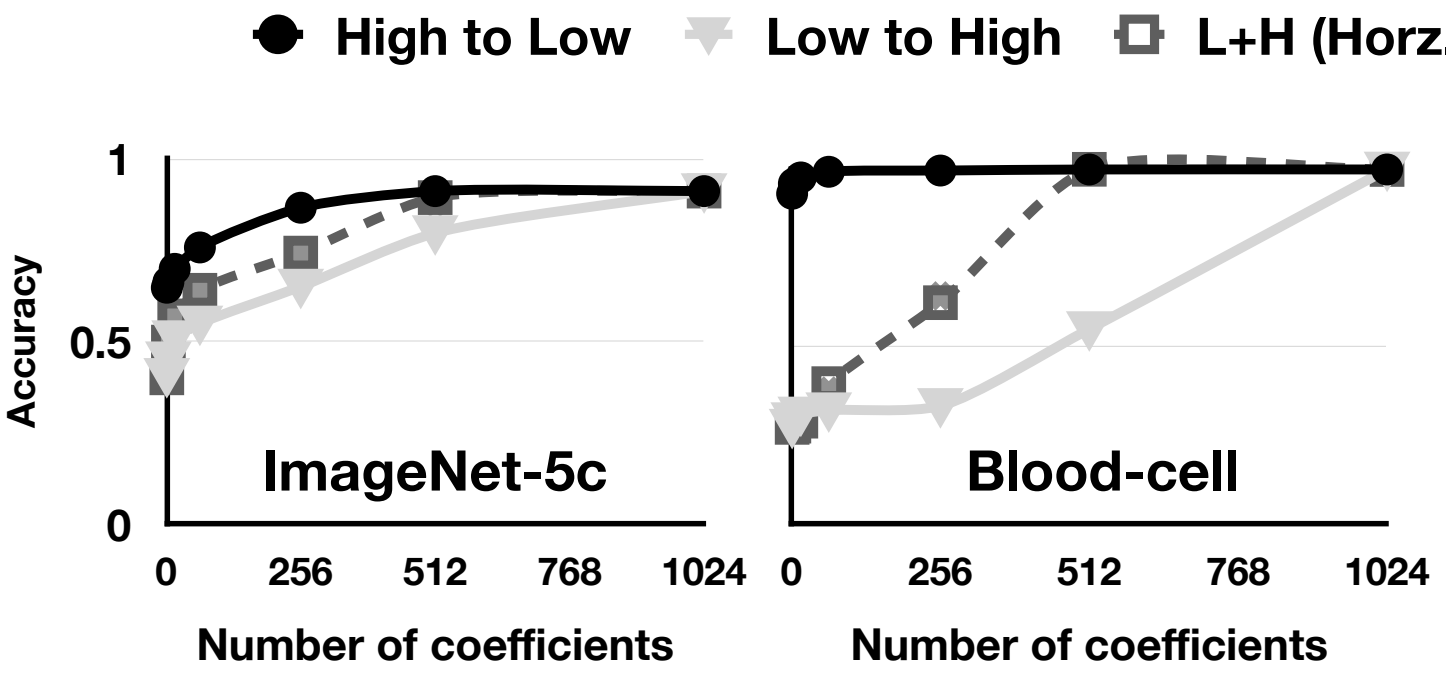


ResNet50, A100, PyTorch v1

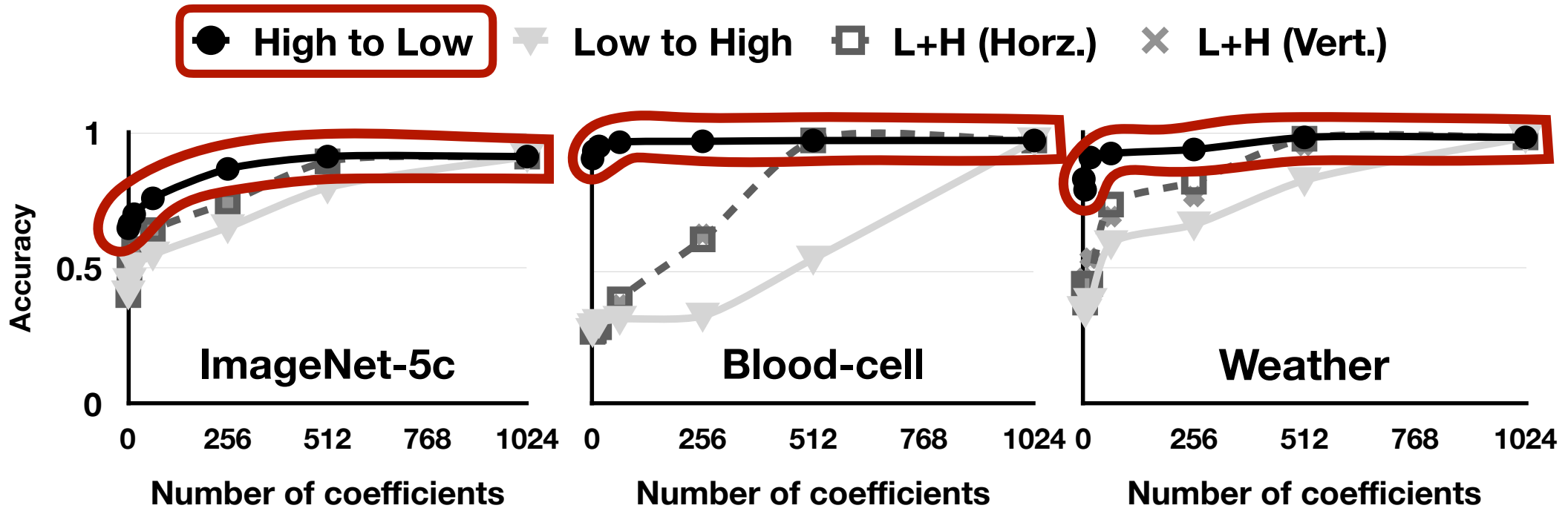


DASlab
@ Harvard SEAS

ResNet50, A100, PyTorch v1



ResNet50, A100, PyTorch v1



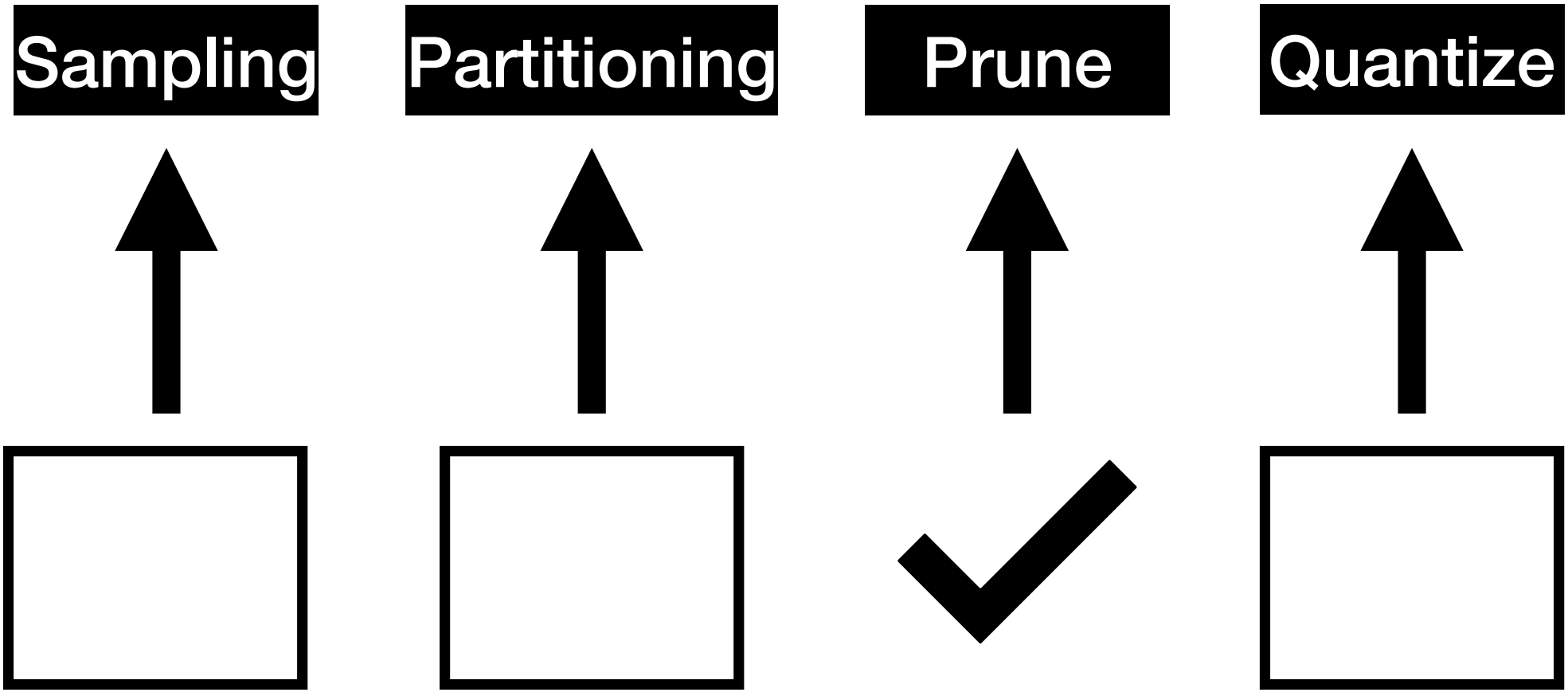
DASlab
@ Harvard SEAS

Pruning is from high to low frequency

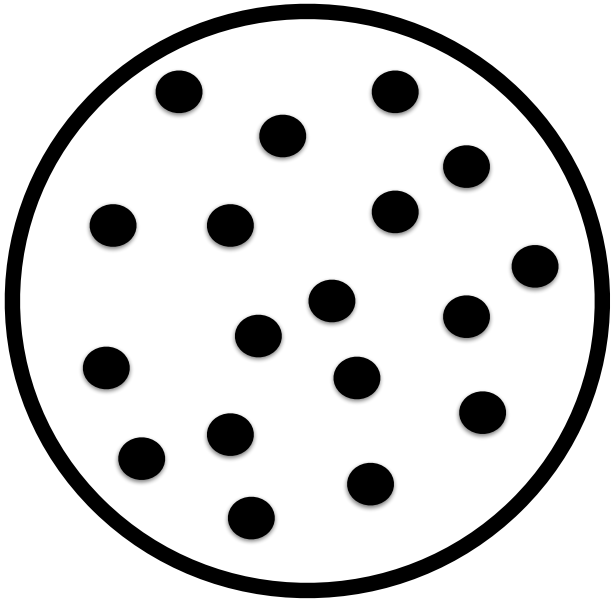
Pruning Domain

$$2^{B \times B} \rightarrow 2B-1$$

Other primitives



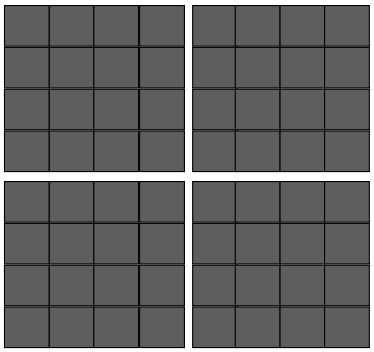
4104 new designs

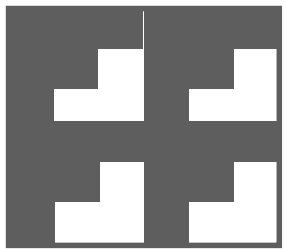


Reading & Writing Algorithms

Sample & Partition → Transform → Prune → Quantize → Encode

Frequency Domain





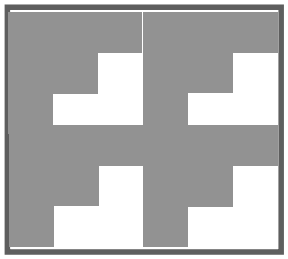
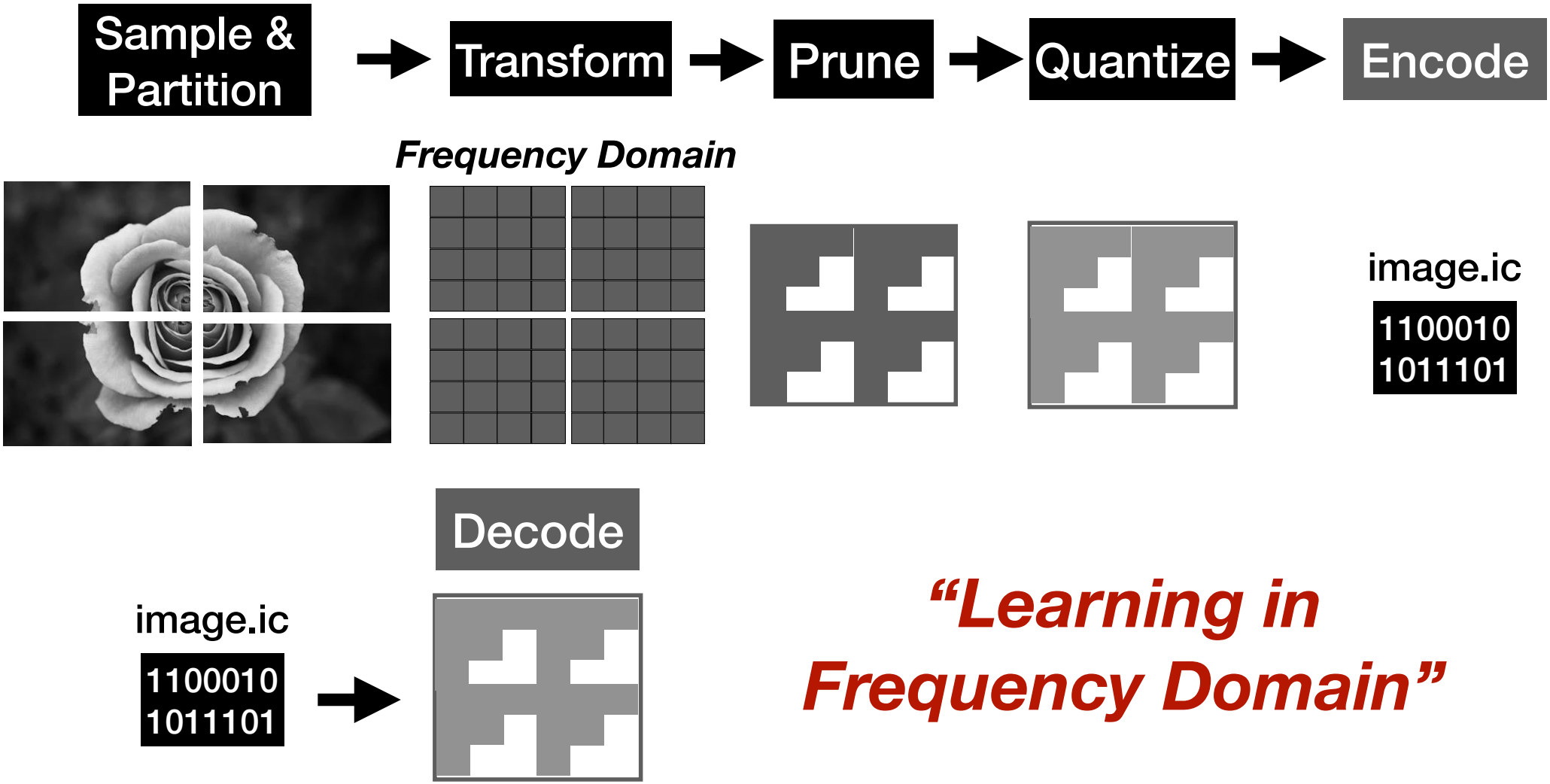


image.ic

1100010
1011101

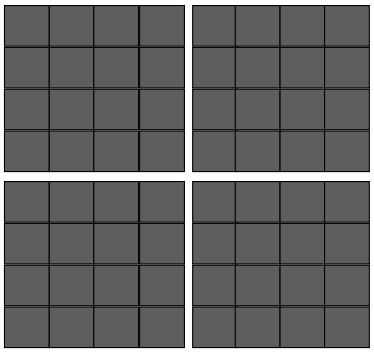


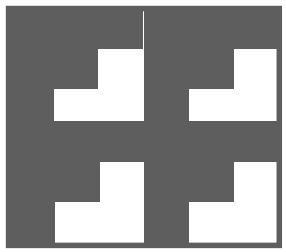
"Learning in Frequency Domain"

Gueguen et al., NIPS'18; Xu et al., CVPR'20; Tancik et al., ICLR'20; Ning et al., ICLR'25

Sample & Partition → Transform → Prune → Quantize → Encode

Frequency Domain





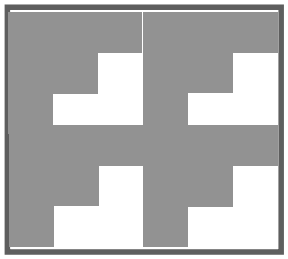


image.ic

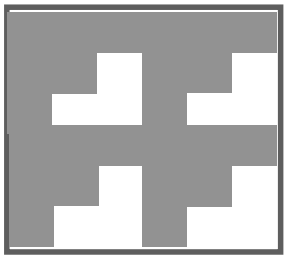
1100010
1011101

Decode

image.ic

1100010
1011101

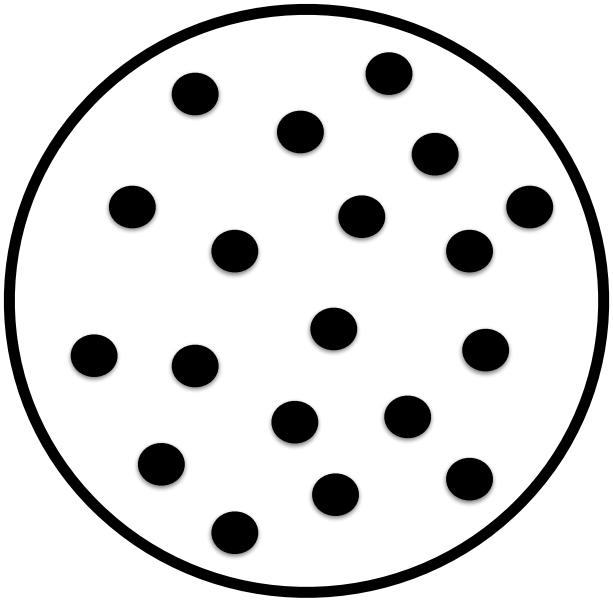
→



✓ **Fast & Simple Reconstruction**

✓ **Low GPU Time**

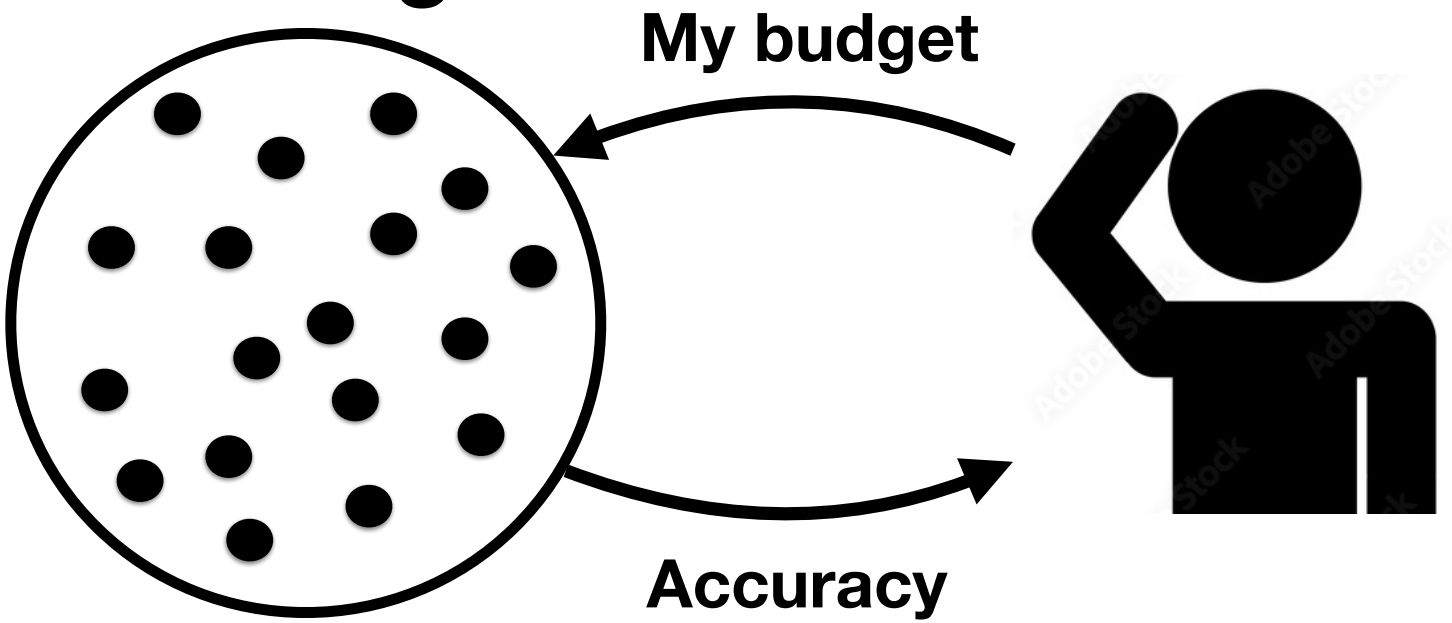
Design space



Search

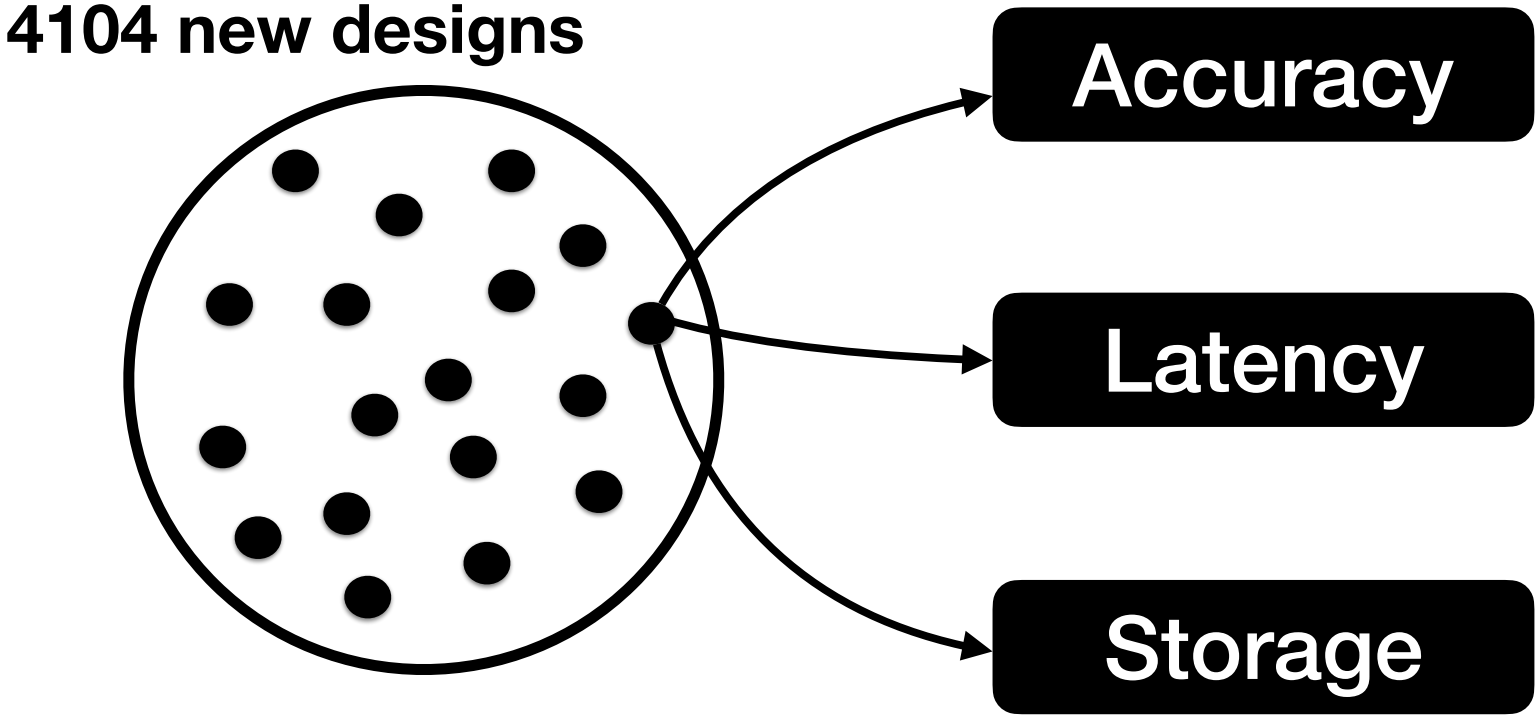
DASlab
@ Harvard SEAS

4104 new designs



DASlab
@ Harvard SEAS

Performance models



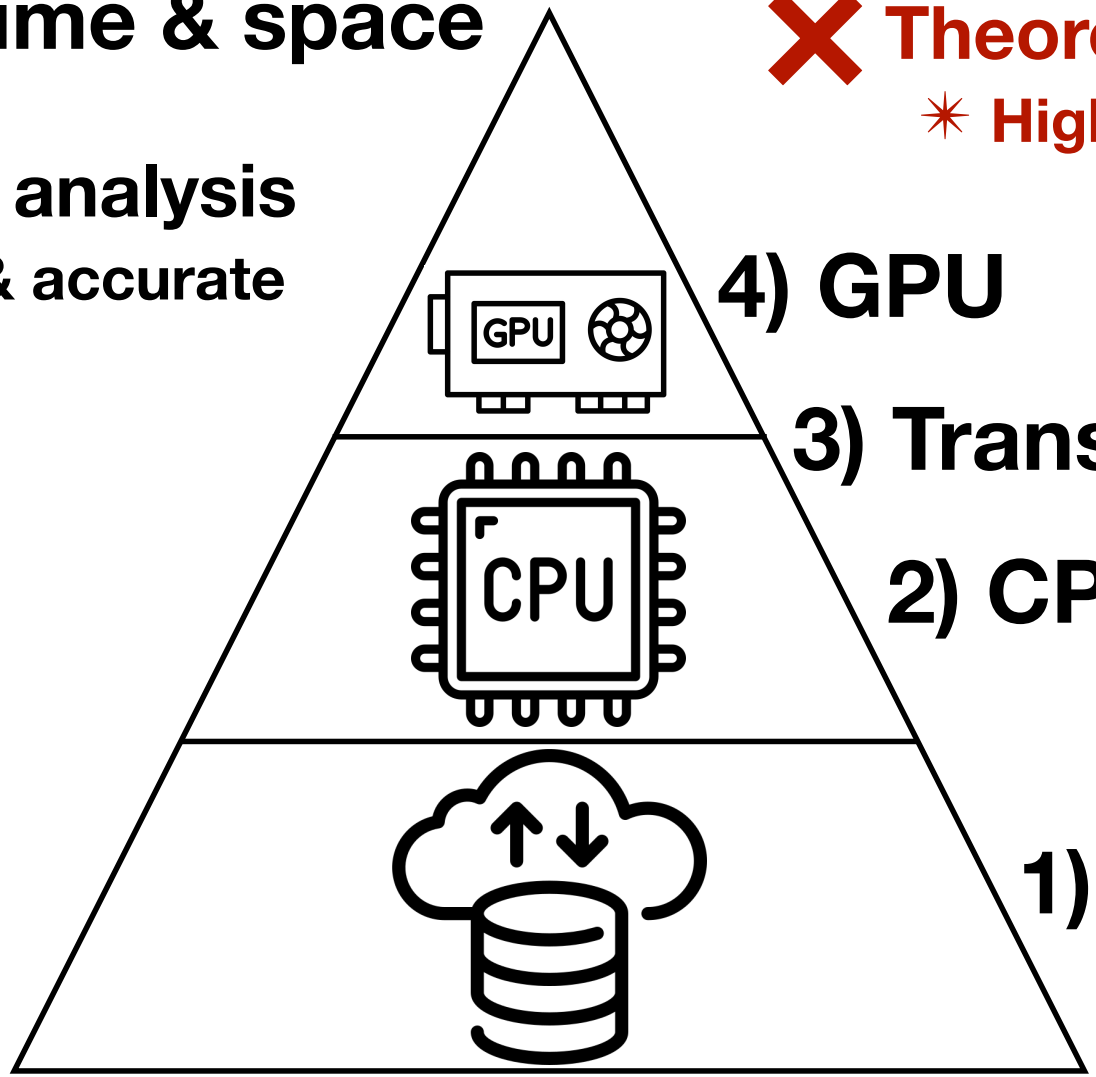
Inf./training time & space

✔ **Empirical analysis**

✳ Cheap & accurate

✖ **Theoretical model**

✳ Highly error-prone



4) GPU

3) Transfer

2) CPU

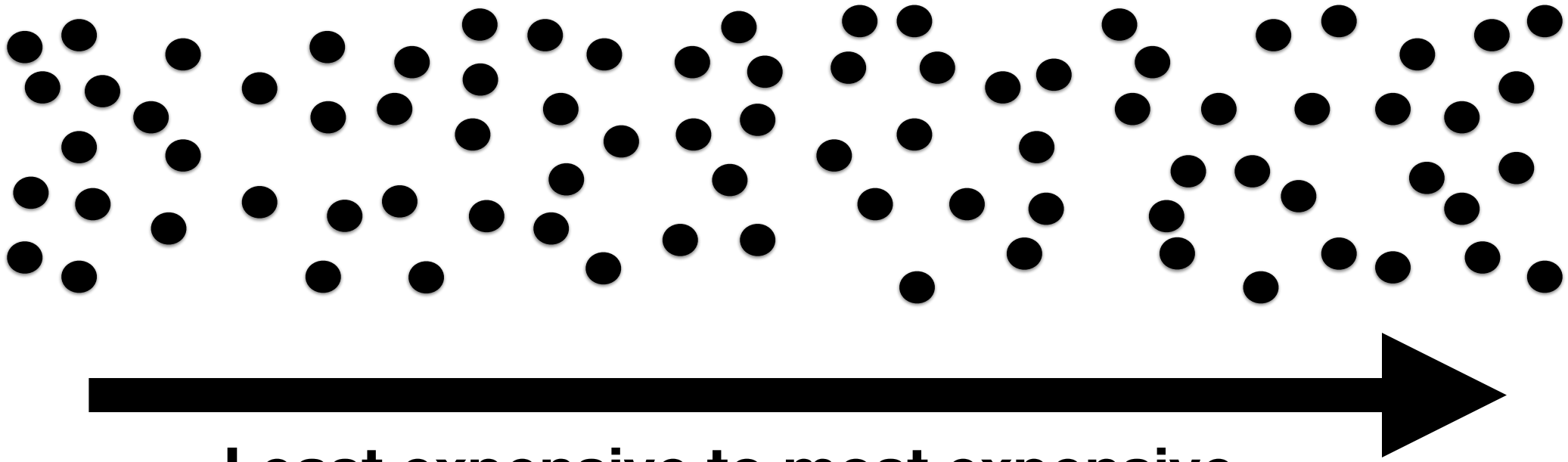
1) Disk I/O

DASlab
@ Harvard SEAS

Accuracy model

Insight 1: Inference Cost and Accuracy are Correlated

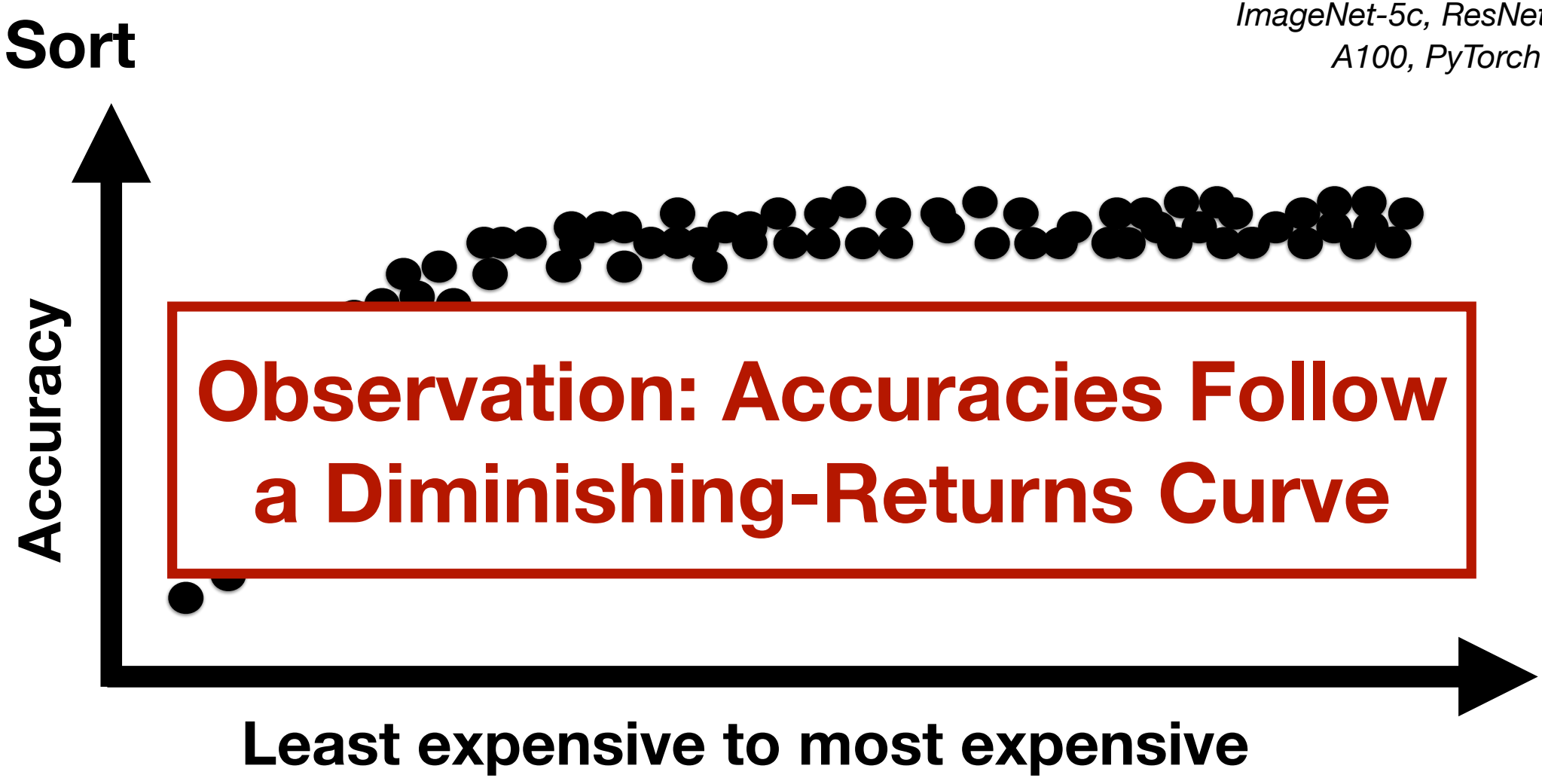
Sort



Least expensive to most expensive

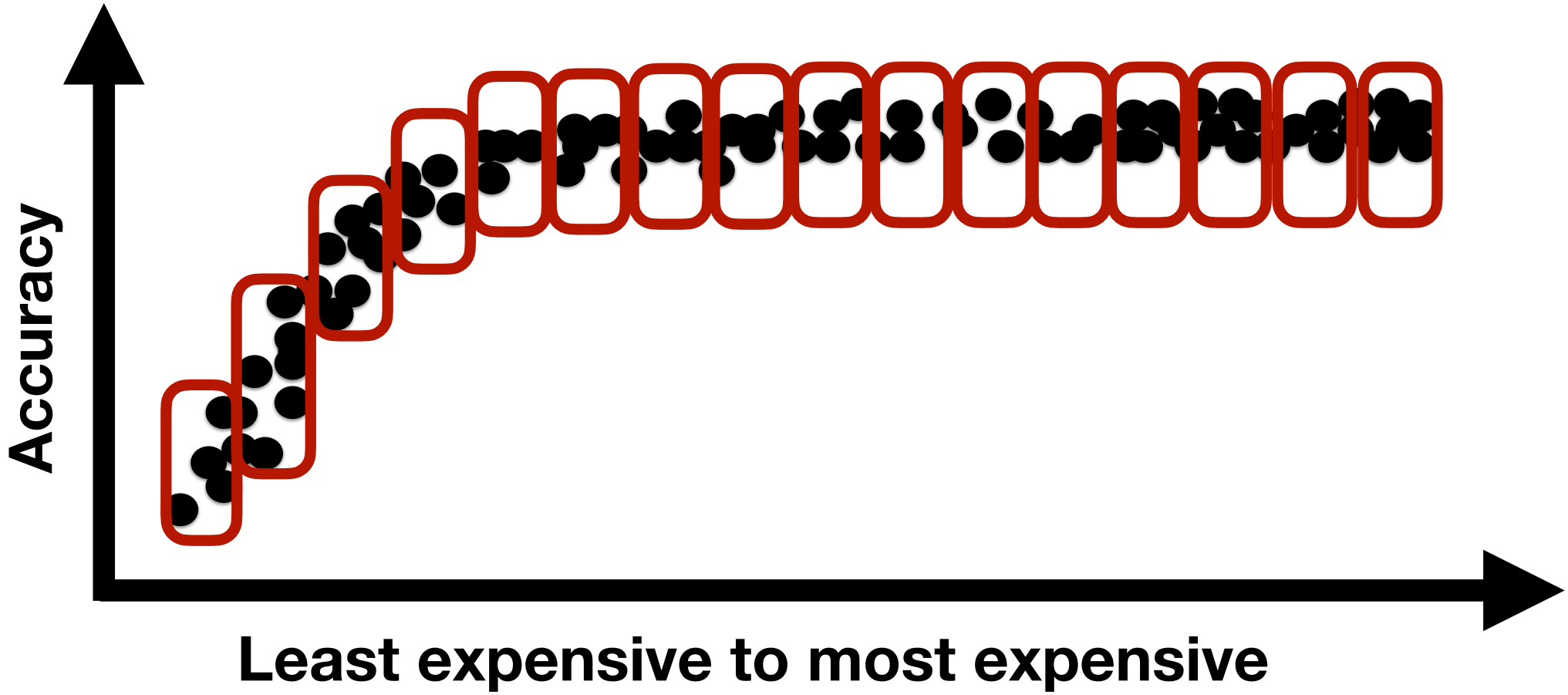
DASlab @ Harvard SEAS

ImageNet-5c, ResNet50
A100, PyTorch v1



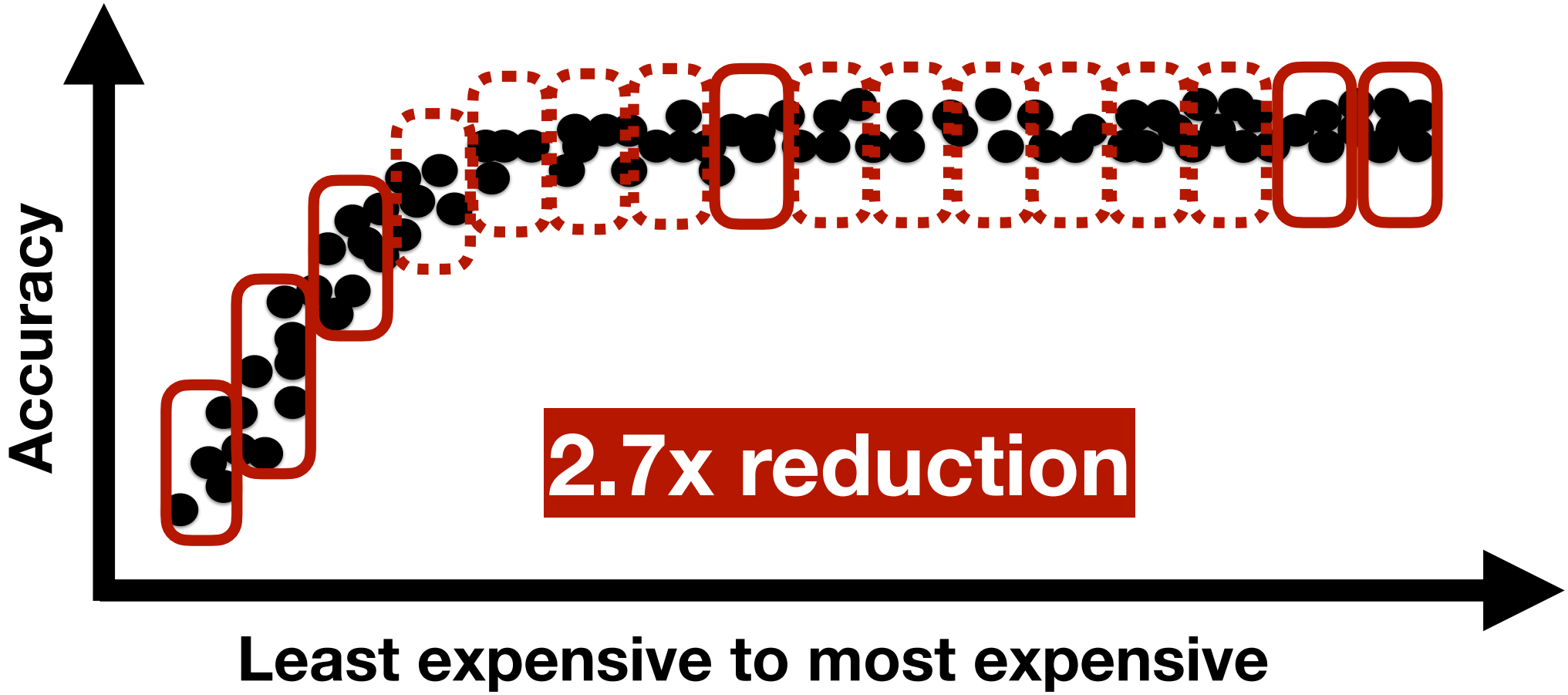
DASlab @ Harvard SEAS

Bucket-sampling & interpolation



DASlab
@ Harvard SEAS

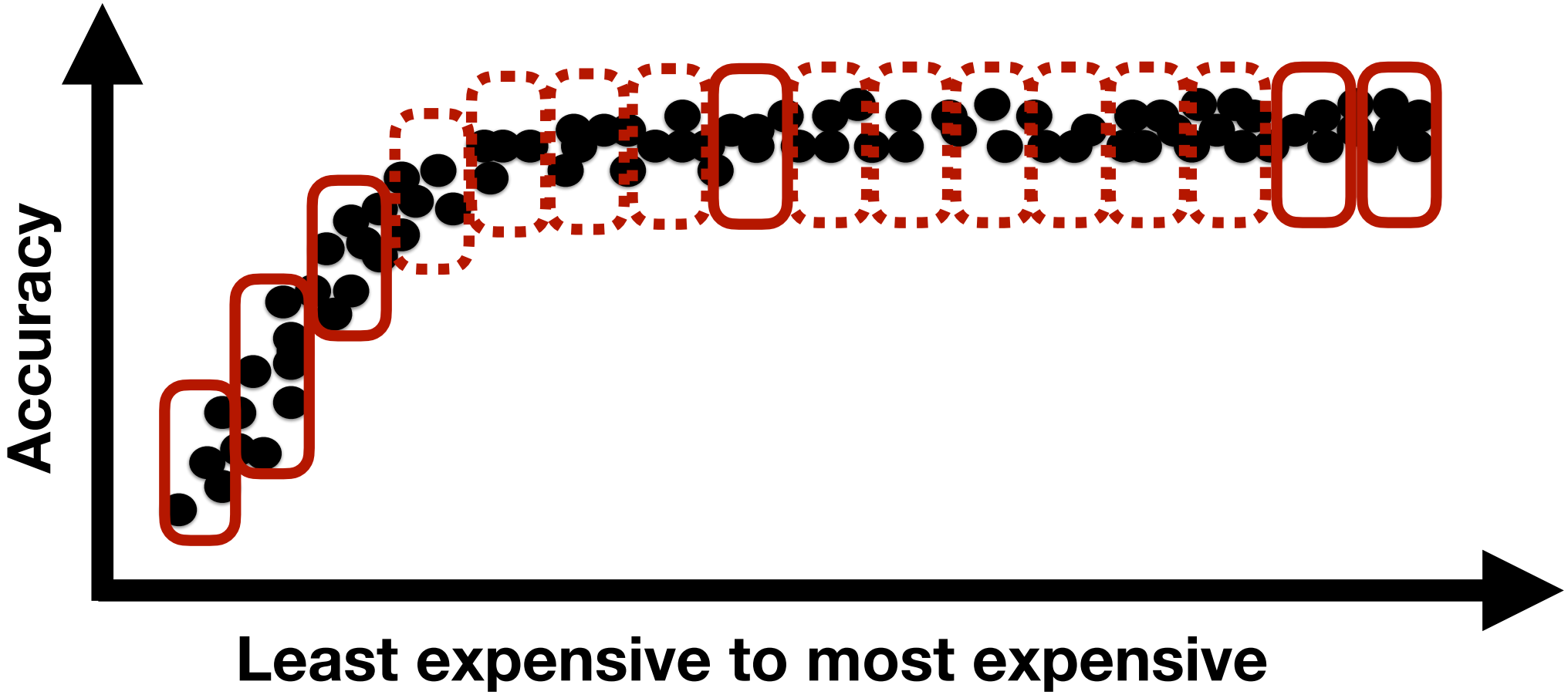
Bucket-sampling & interpolation



DASlab
@ Harvard SEAS

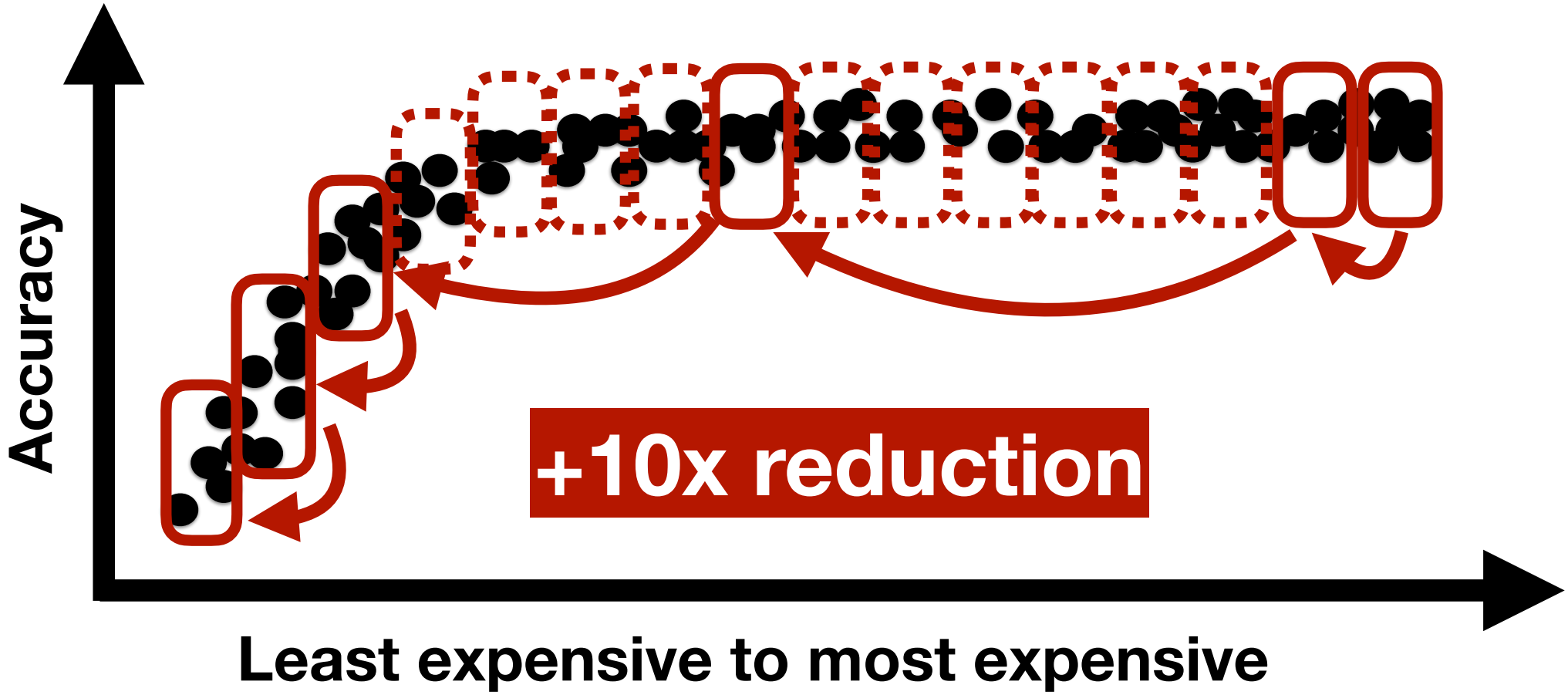
Insight 2: ***Storage Formats are Correlated***

Bucket-sampling & interpolation



DASlab
@ Harvard SEAS

Transfer learning

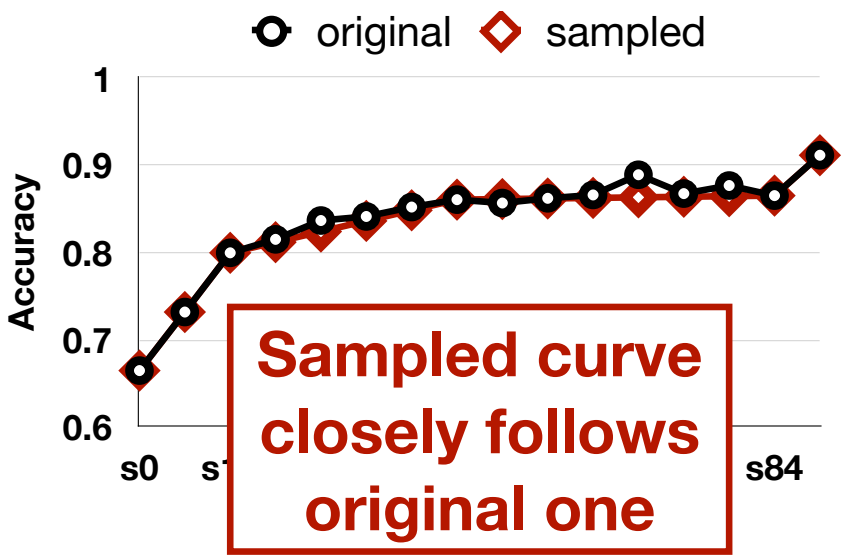


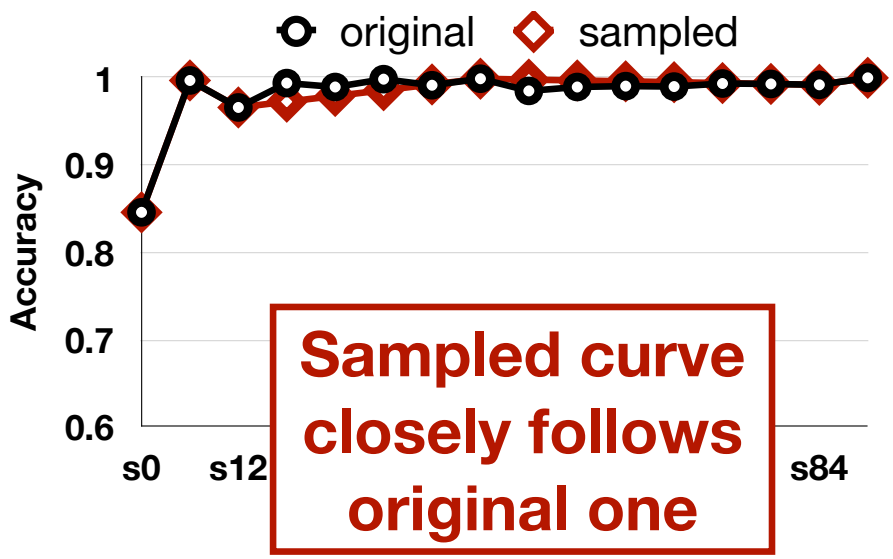
DASlab
@ Harvard SEAS

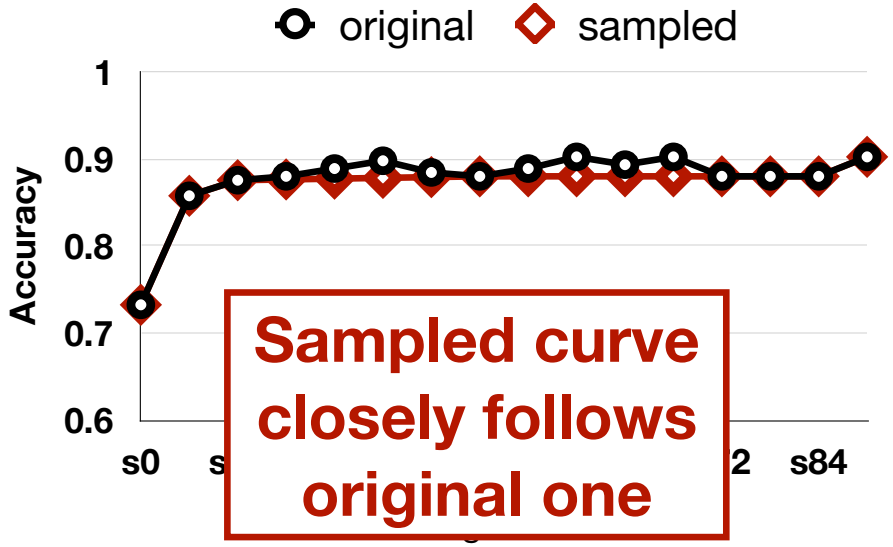
~4K AI model trainings → **150 AI model trainings**

Verification

ResNet50, A100, PyTorch v1

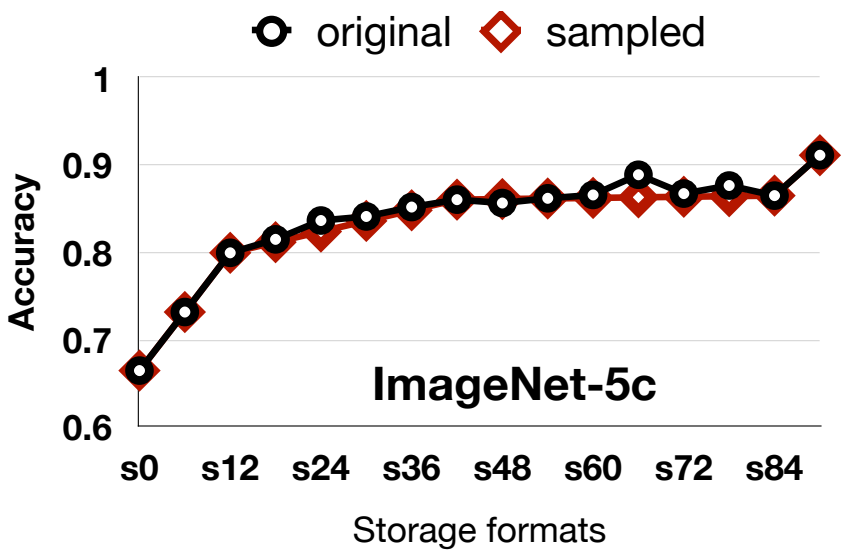


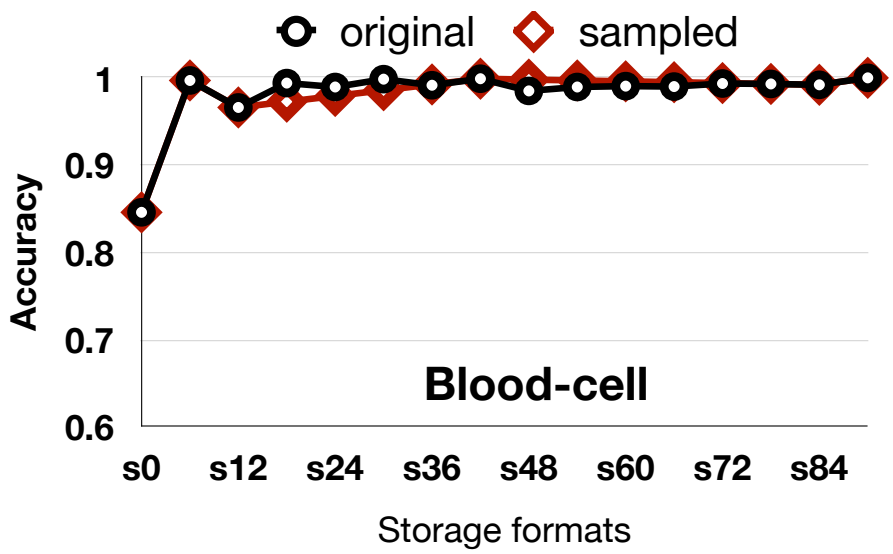




DASlab @ Harvard SEAS

ResNet50, A100, PyTorch v1

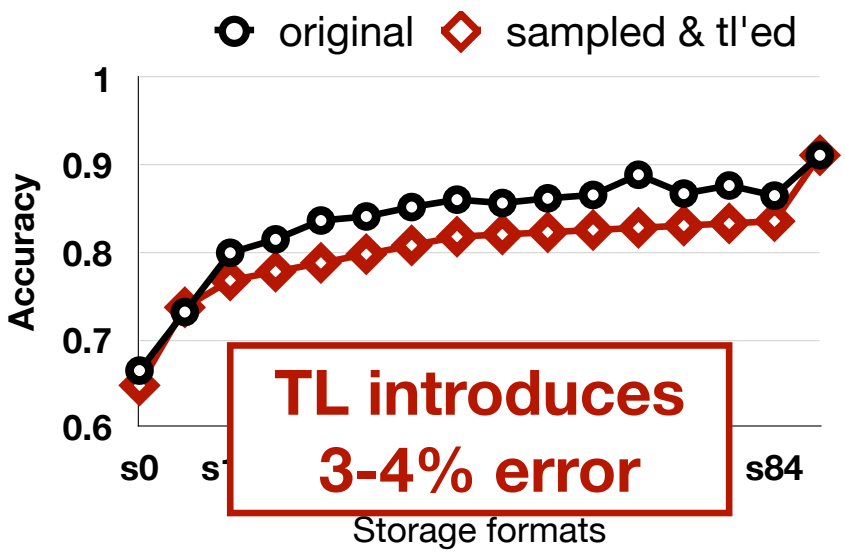


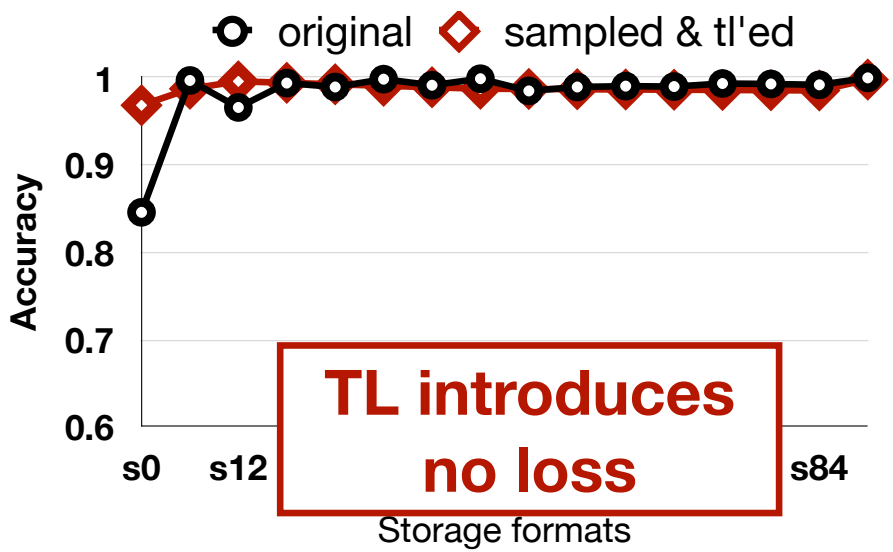


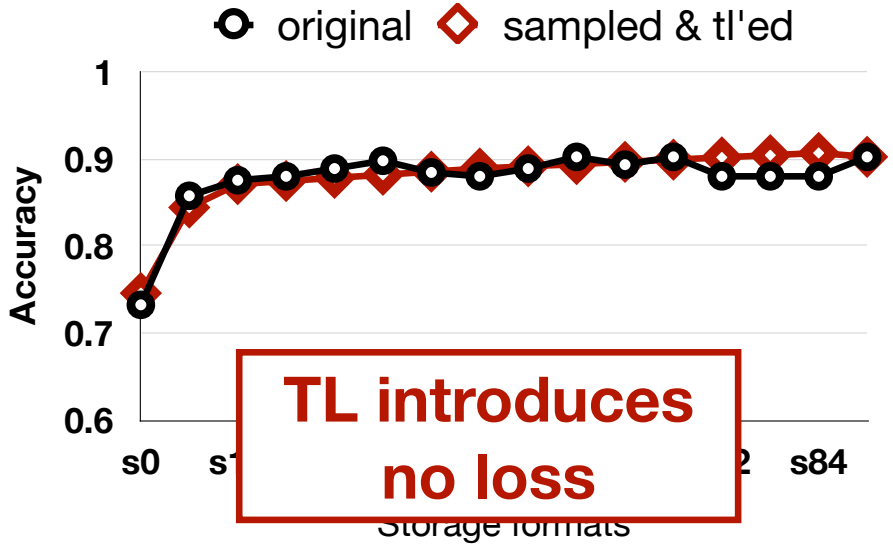
Sampling reduces cost with no loss

DASlab @ Harvard SEAS

ResNet50, A100, PyTorch v1

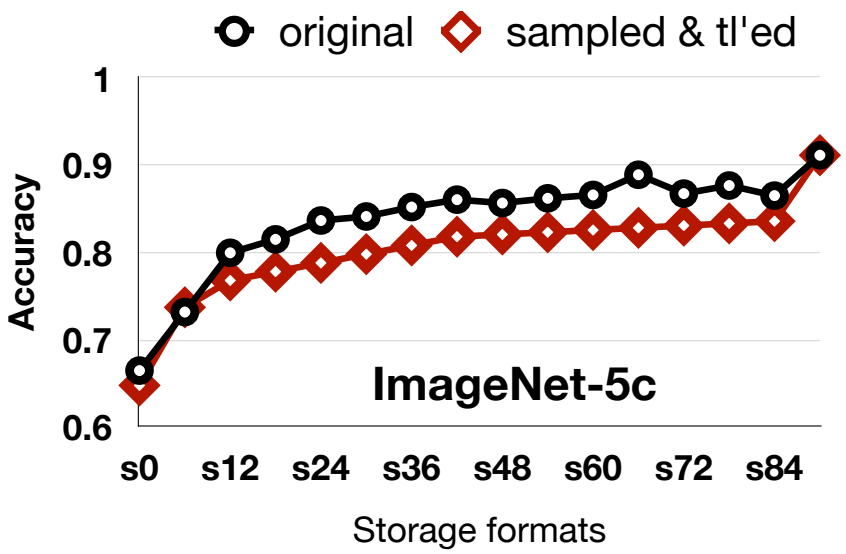


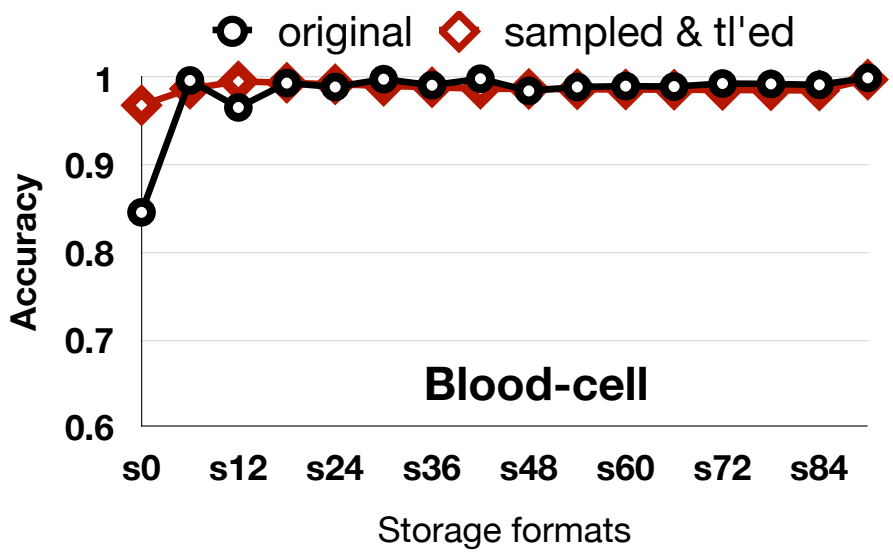


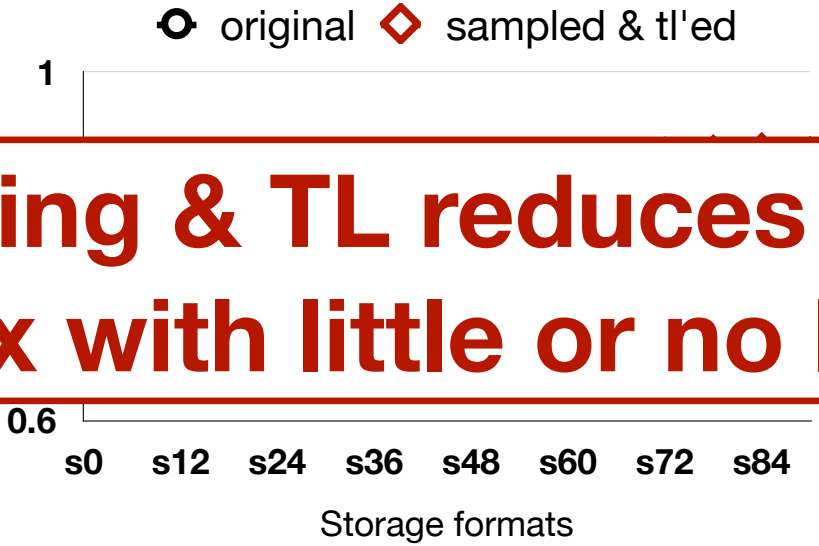


DASlab @ Harvard SEAS

ResNet50, A100, PyTorch v1



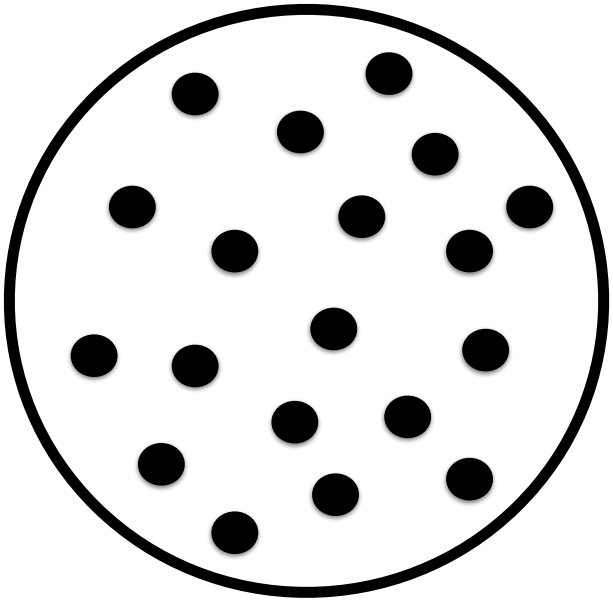




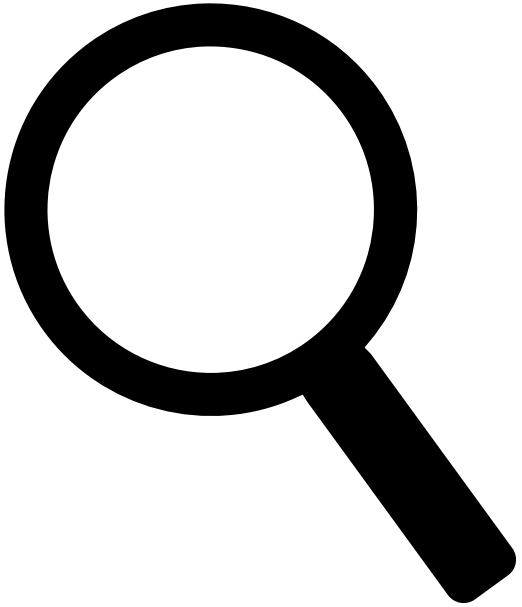
Sampling & TL reduces cost by 30x with little or no loss

DASlab @ Harvard SEAS

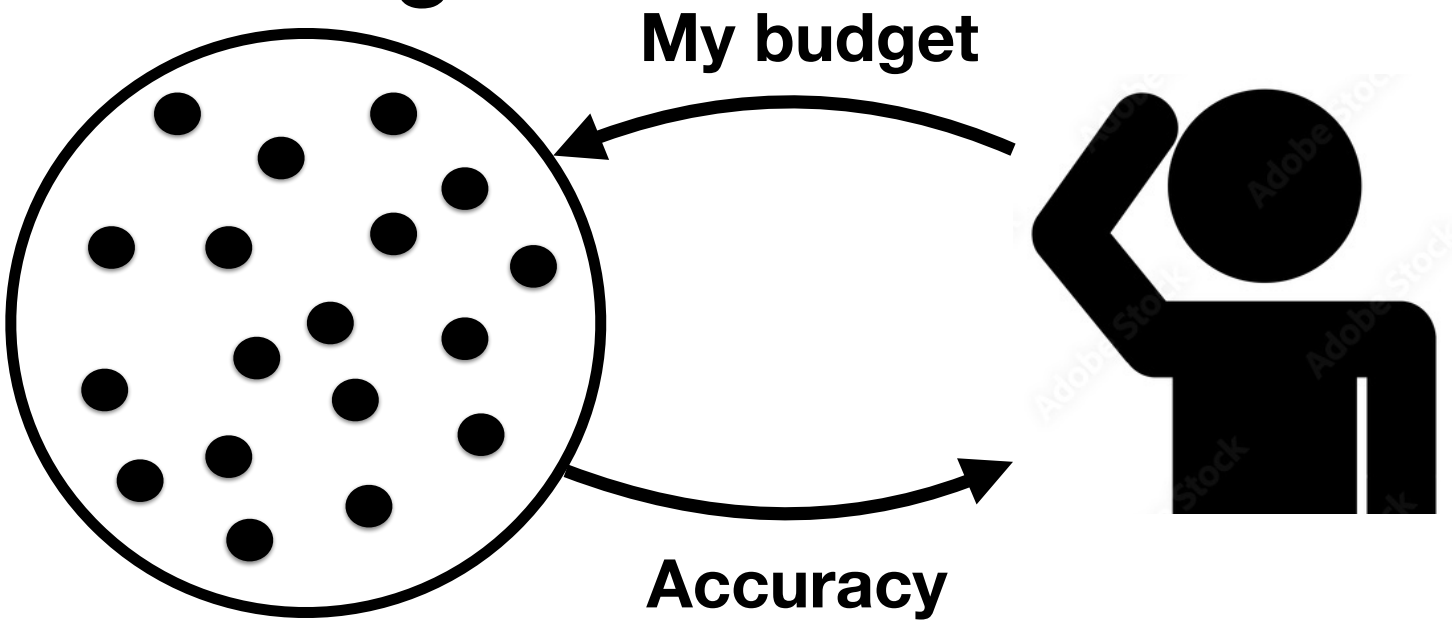
Design space



Search



4104 new designs



DASlab
@ Harvard SEAS

Machine: Nvidia A100 w/ 4 GPUs

Datasets:

- 5-class ImageNet
- Blood-cell images
- Weather prediction
- 50-class + full ImageNet

AI Models: ResNet, EfficientNet, MobileNet

Hyper-parameters: Standard from PyTorch

Baseline: JPEG

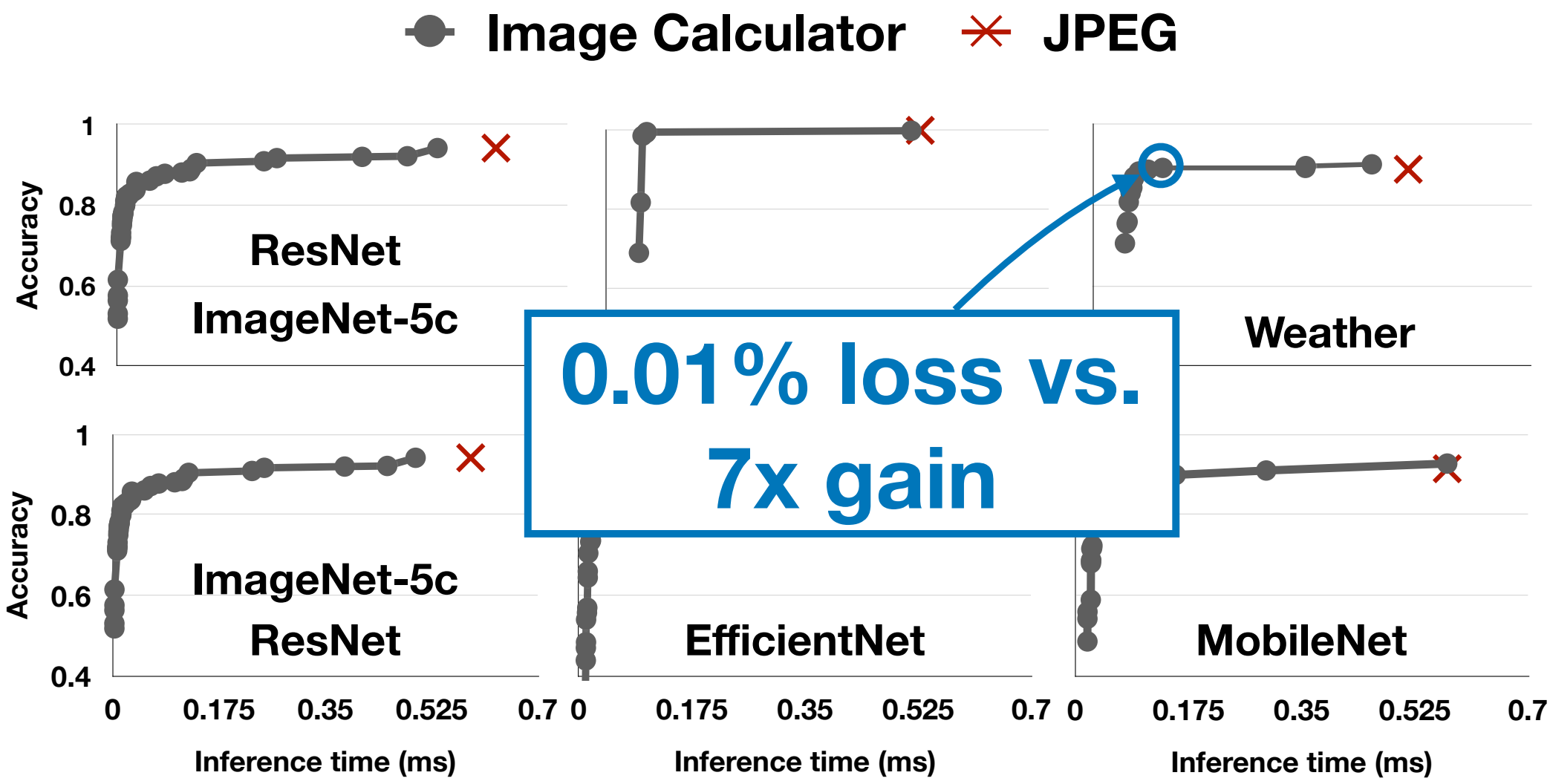


Image Calculator
JPEG
Accuracy
1
0.8
0.6
0.4
ResNet
ImageNet-5c
Weather
0.01% loss vs.
7x gain
ImageNet-5c
ResNet
EfficientNet
MobileNet
0
0.175
0.35
0.525
0.7
Inference time (ms)

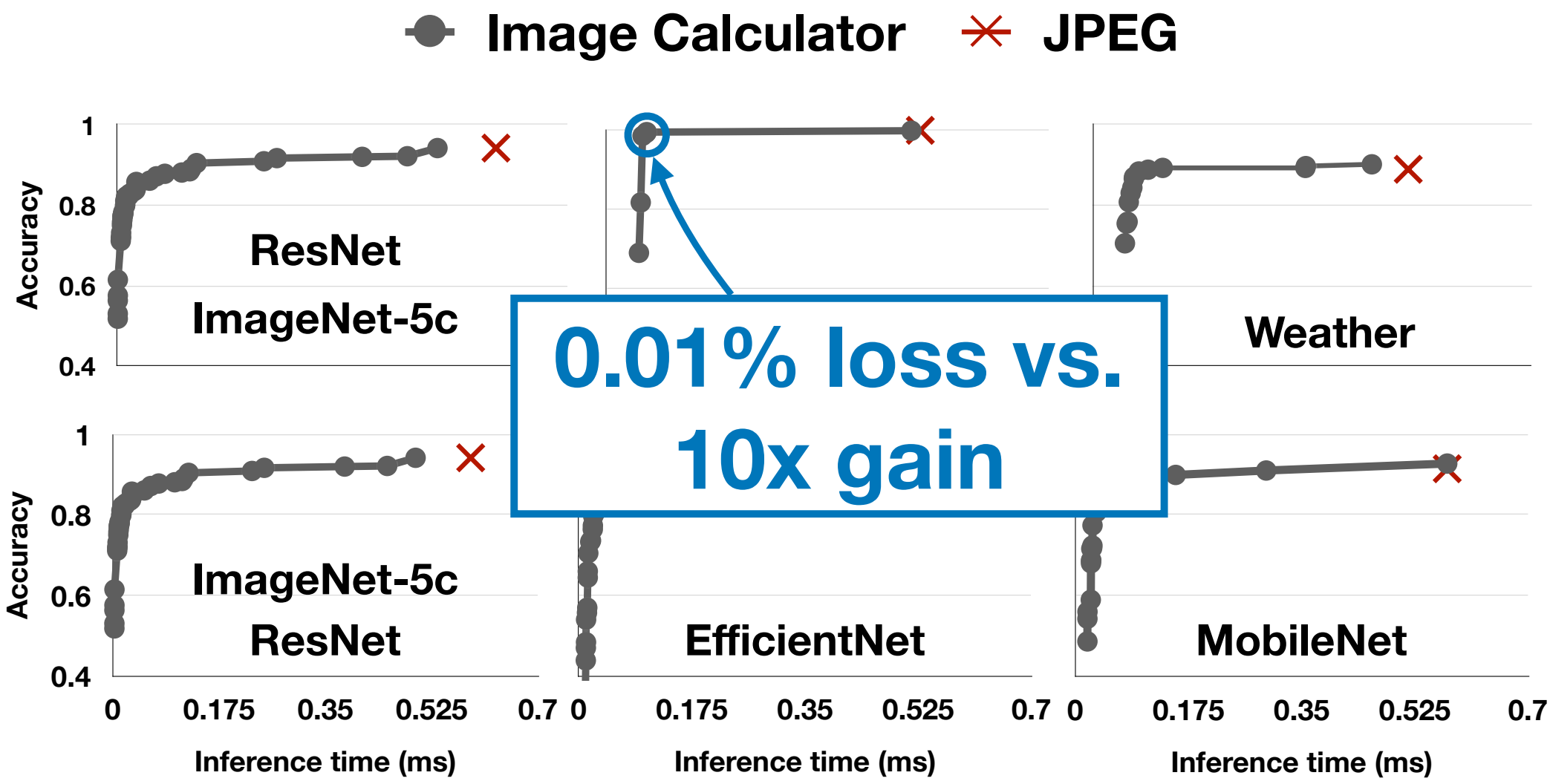


Image Calculator
JPEG
Accuracy
1
0.8
0.6
0.4
ResNet
ImageNet-5c
Weather
0.01% loss vs.
10x gain
ImageNet-5c
ResNet
EfficientNet
MobileNet
0
0.175
0.35
0.525
0.7
Inference time (ms)

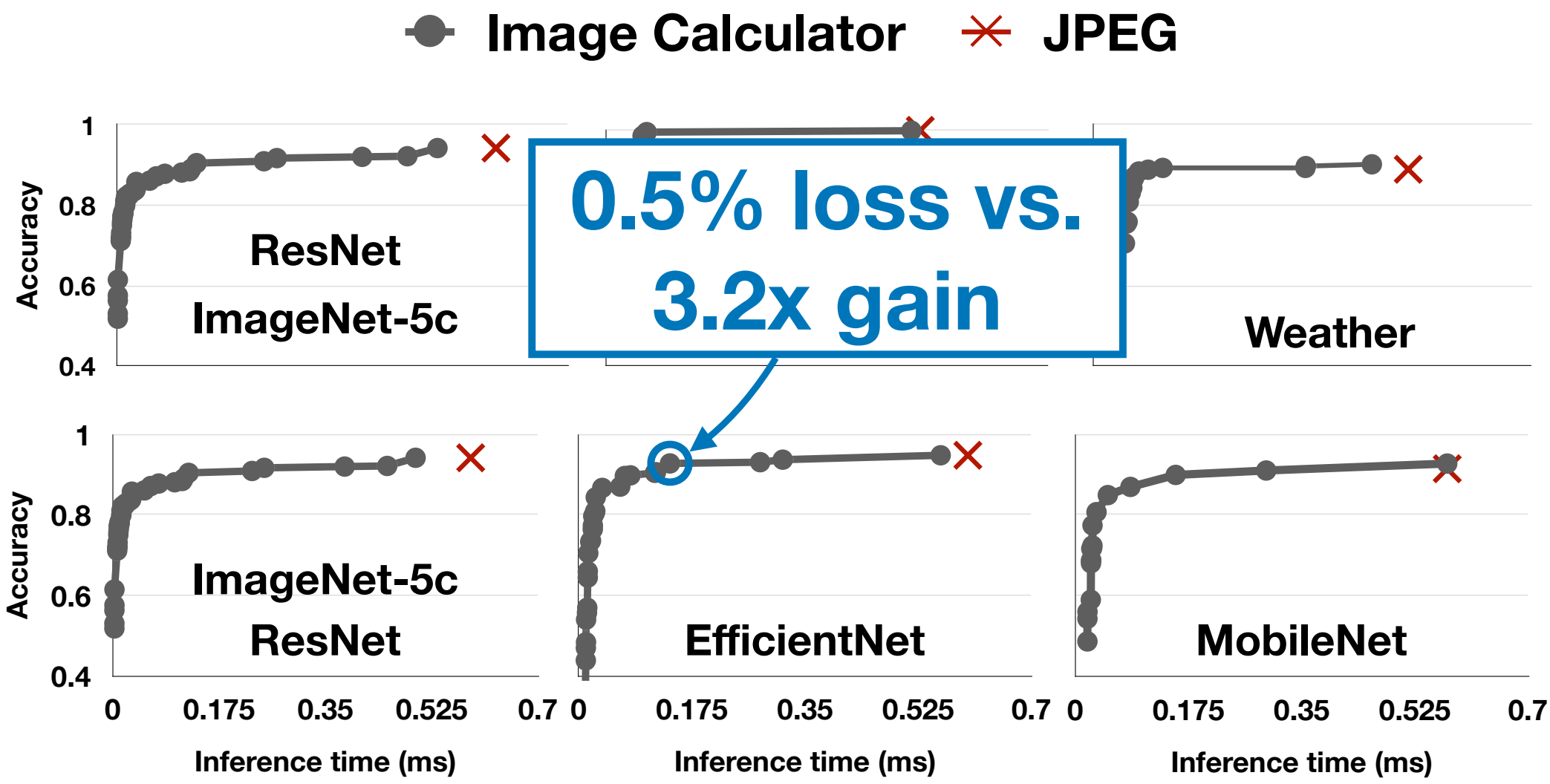
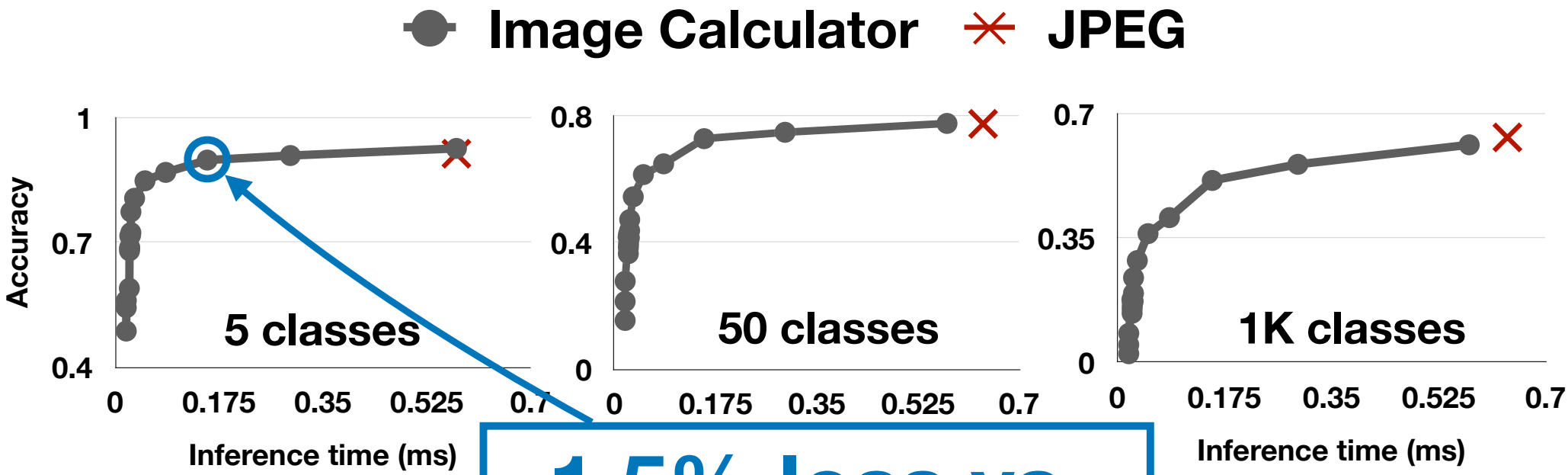


Image Calculator
JPEG
Accuracy
1
0.8
0.6
0.4
ResNet
ImageNet-5c
0.5% loss vs.
3.2x gain
Weather
ImageNet-5c
ResNet
EfficientNet
MobileNet
0
0.175
0.35
0.525
0.7
Inference time (ms)

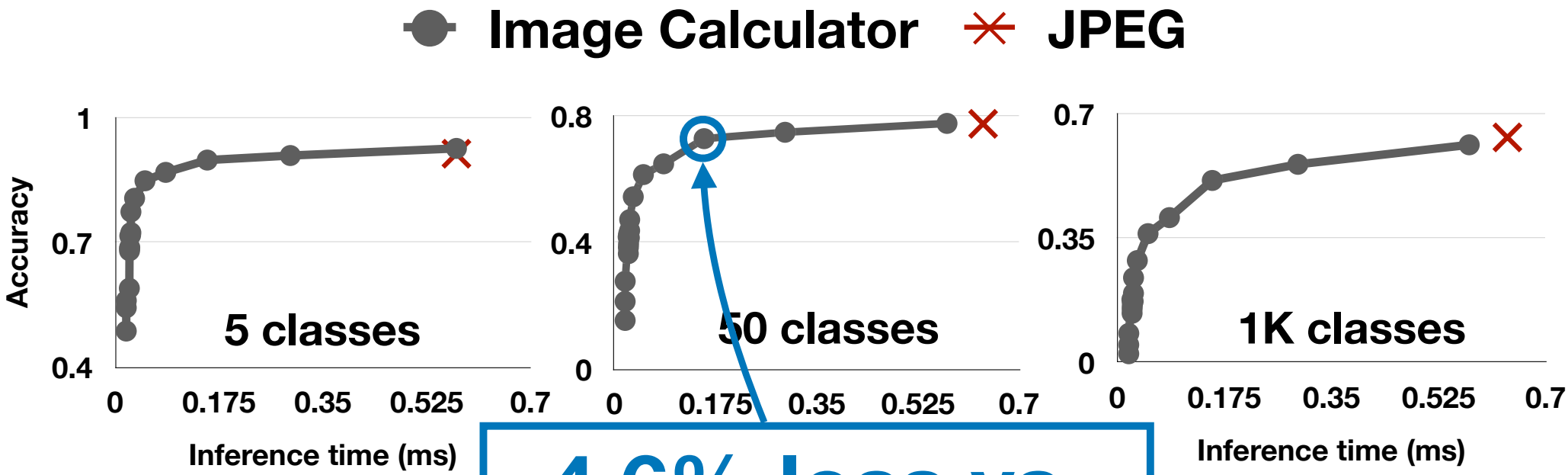
AI Model: MobileNet



1.5% loss vs.
3.7x gain

DASlab
@ Harvard SEAS

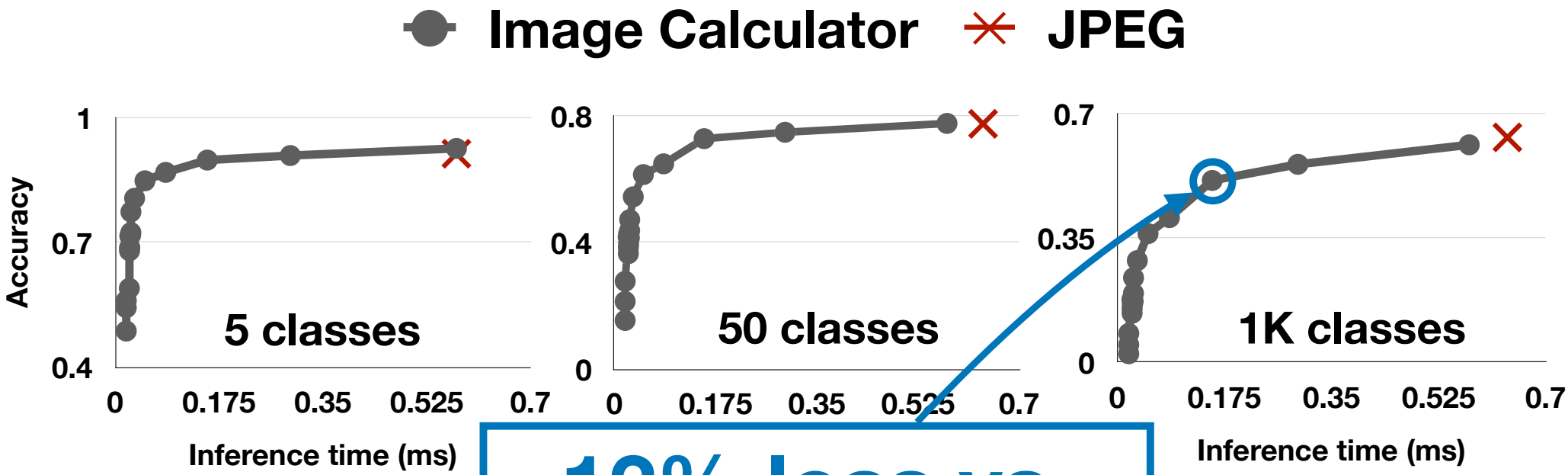
AI Model: MobileNet



4.6% loss vs.
3.7x gain

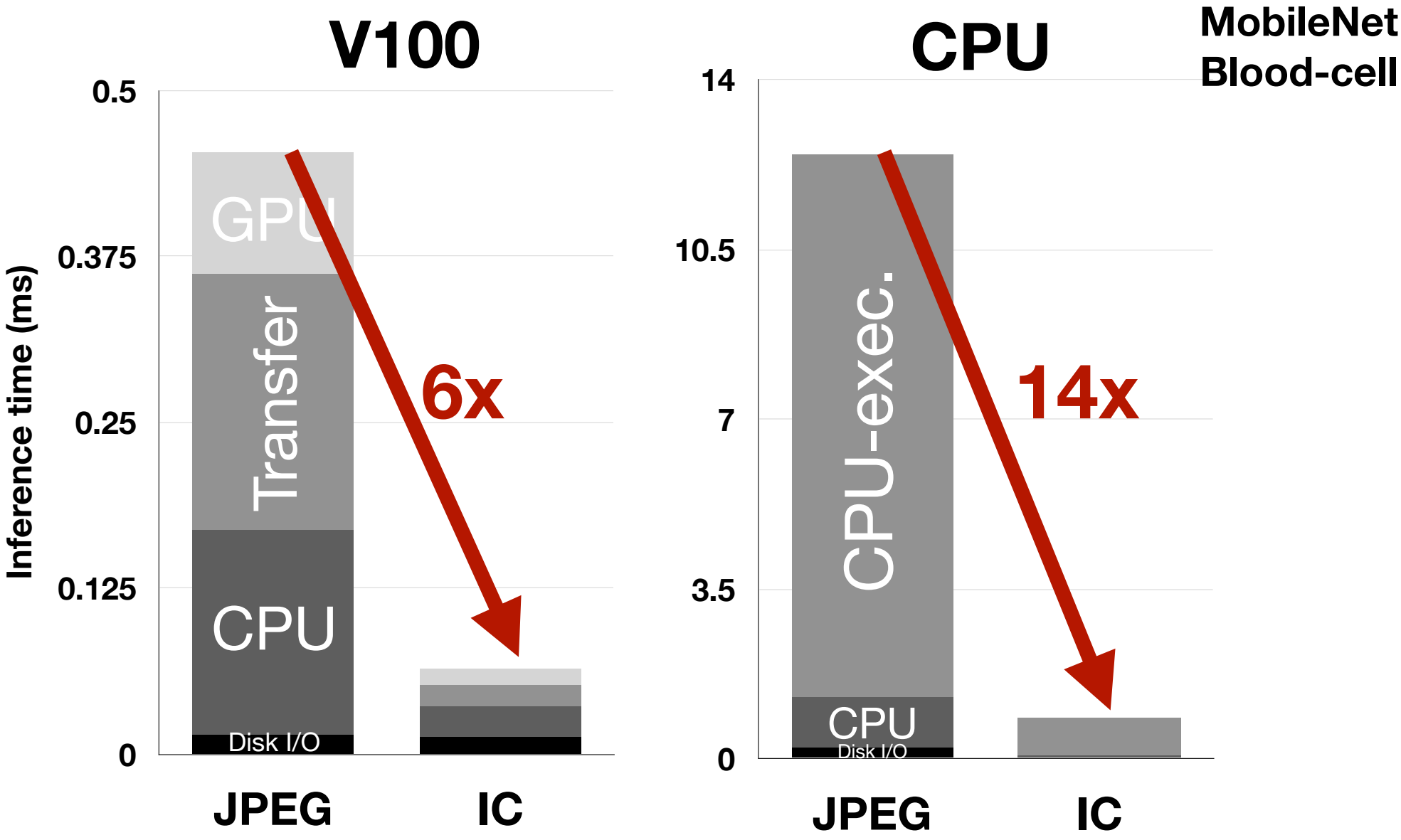
DASlab
@ Harvard SEAS

AI Model: MobileNet



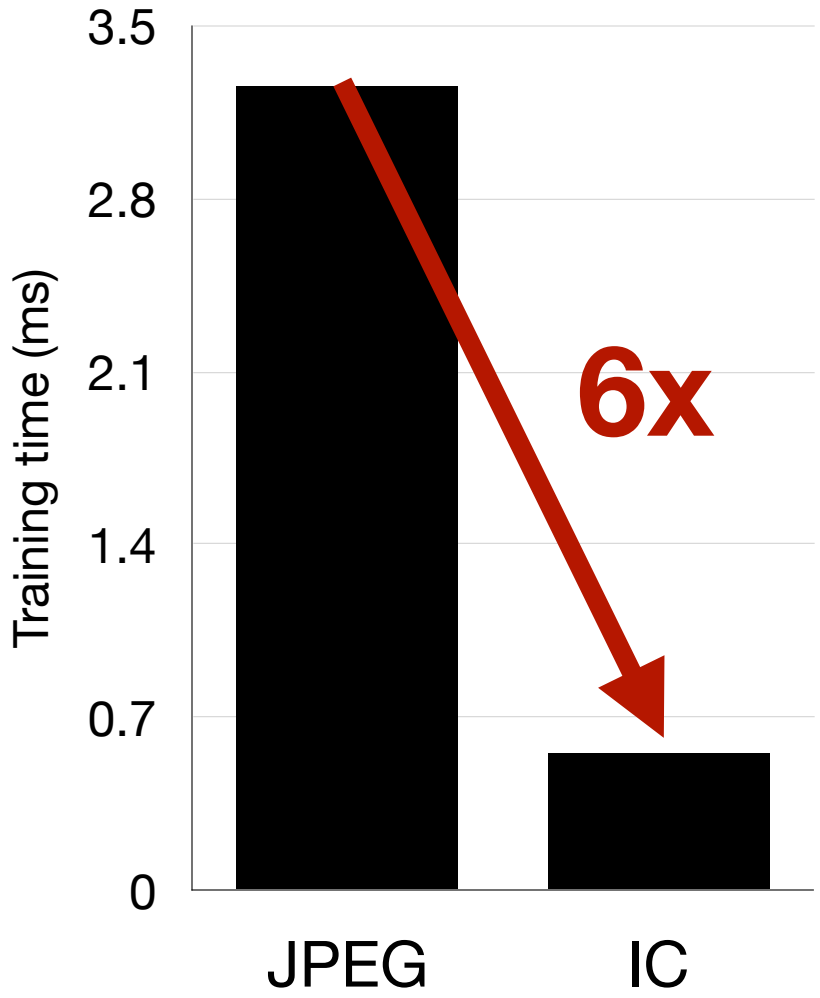
12% loss vs.
3.7x gain

DASlab
@ Harvard SEAS


V100
CPU
MobileNet
Blood-cell
Inference time (ms)
0.5
0.375
0.25
0.125
0
GPU
Transfer
CPU
Disk I/O
6x
14
10.5
7
3.5
0
CPU-exec.
CPU
Disk I/O
14x
JPEG
IC
JPEG
IC

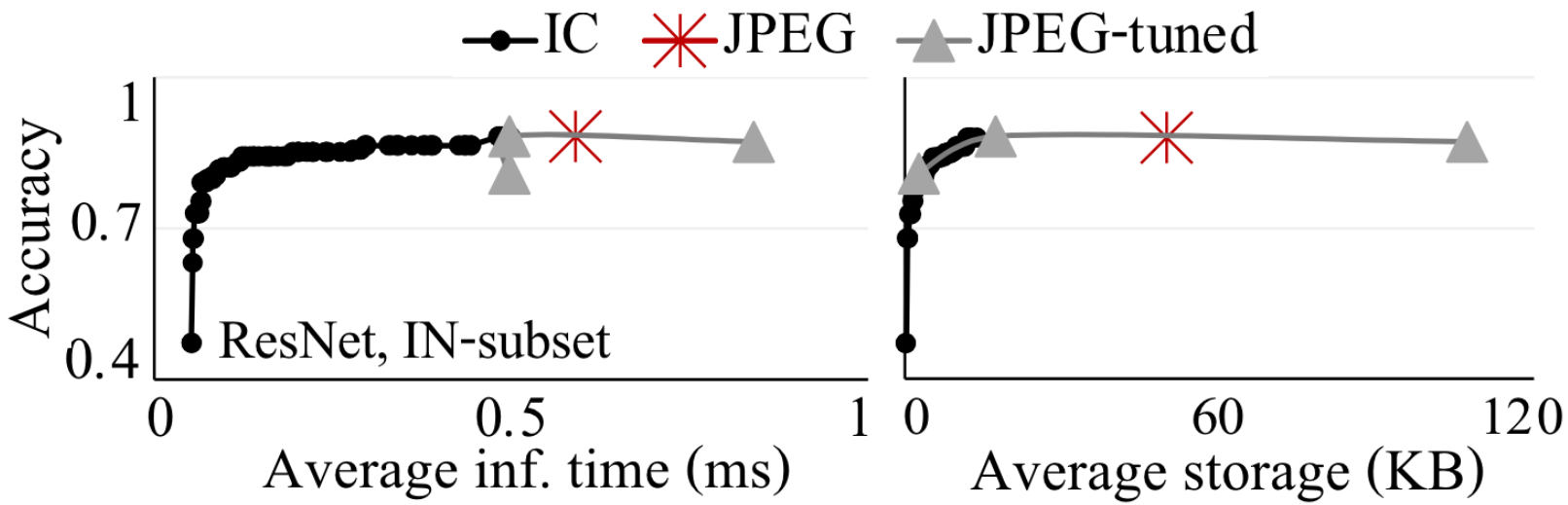
IC reduces training time

Blood-cell, ResNet50
A100, PyTorch v1

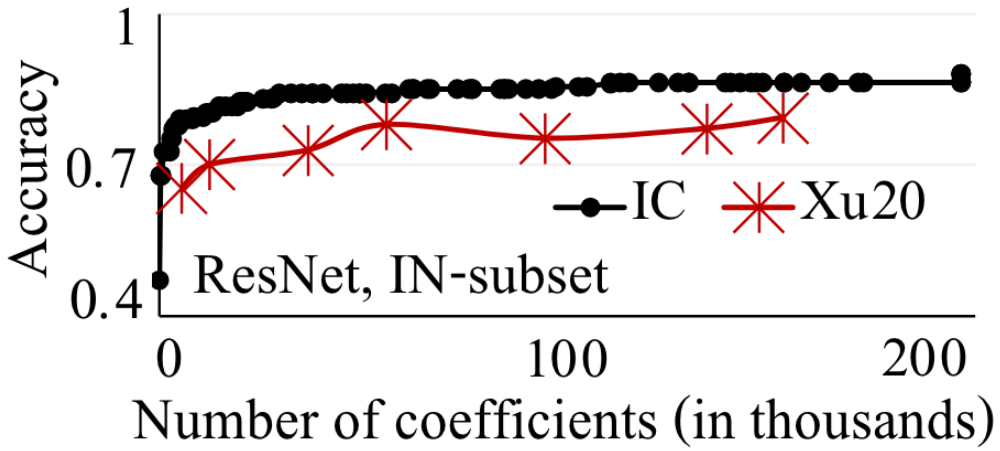


DASlab
@ Harvard SEAS

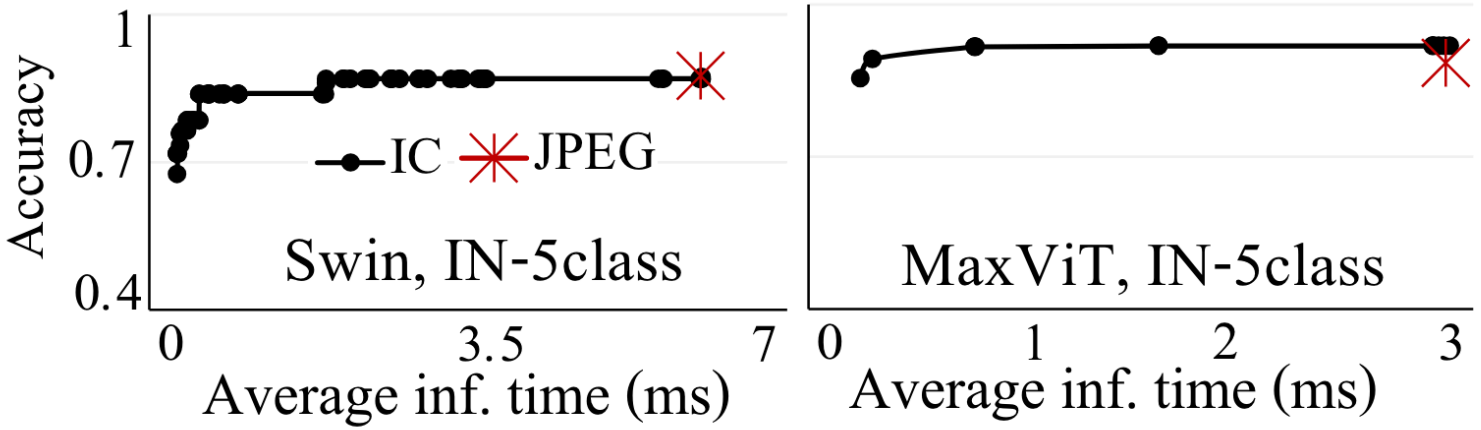
JPEG vs. JPEG-tuned



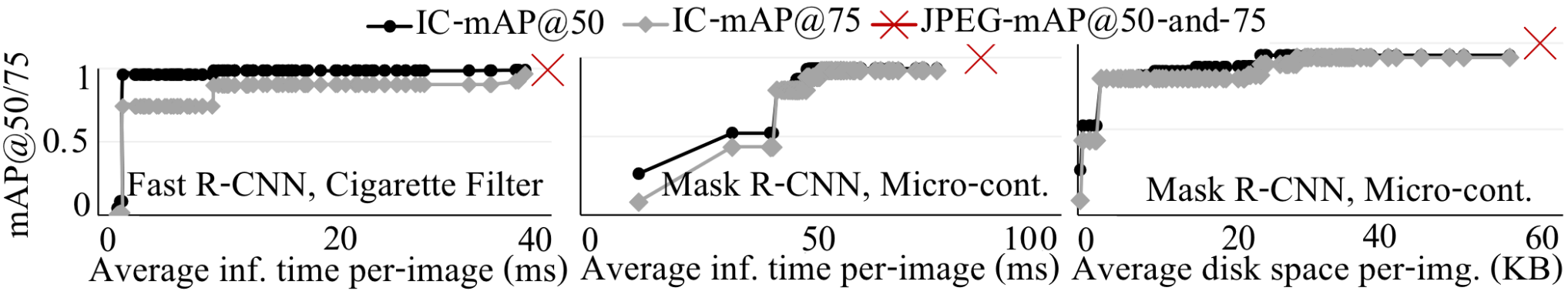
Related work, “Learned JPEG” (CVPR’20)



Visual Transformers



Object detection



Latency-accuracy **Latency-accuracy** **Storage-accuracy**

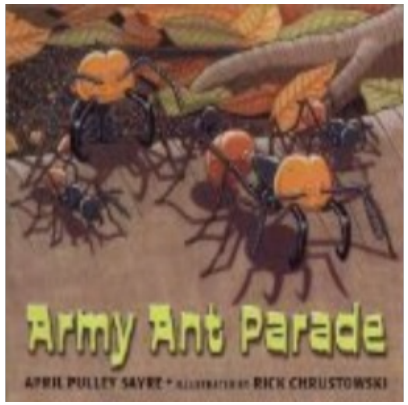
DASlab
@ Harvard SEAS

Object detection & instance segmentation — examples

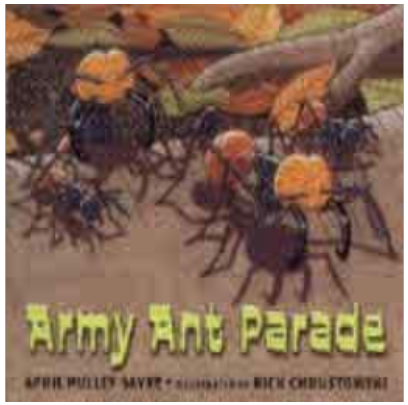
DASlab
@ Harvard SEAS

IC maintains high visual quality

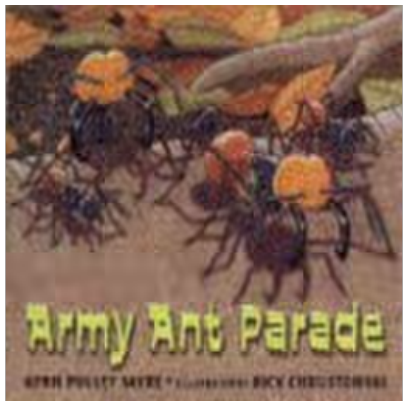
ImageNet-5c
PyTorch v1



Original



60x



100x

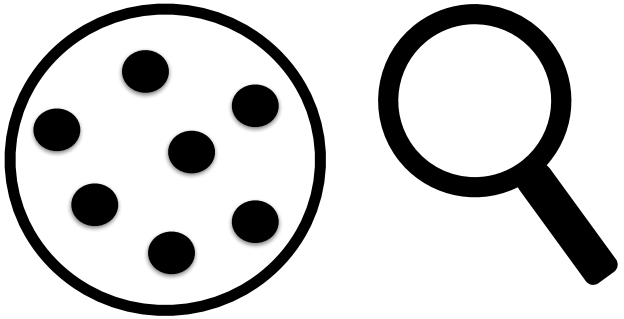
200x

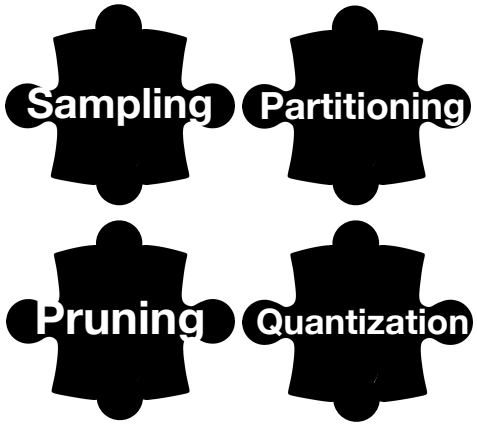
500x

1200x

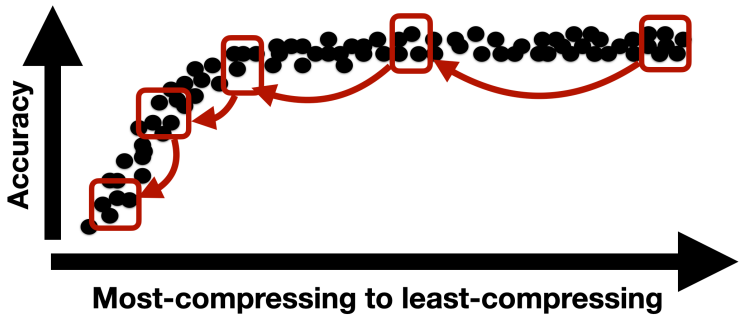
DASlab
@ Harvard SEAS

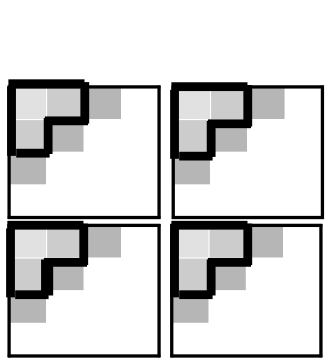
Self-Designing AI Storage

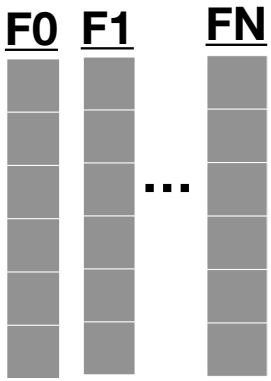




10^{150K} ➡ 4104





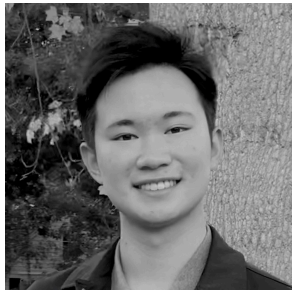


Efficiency vs. Scalability

DASlab @ Harvard SEAS








VE RI TAS

DASlab
@ Harvard SEAS

daslab.seas.harvard.edu

THANKS!