

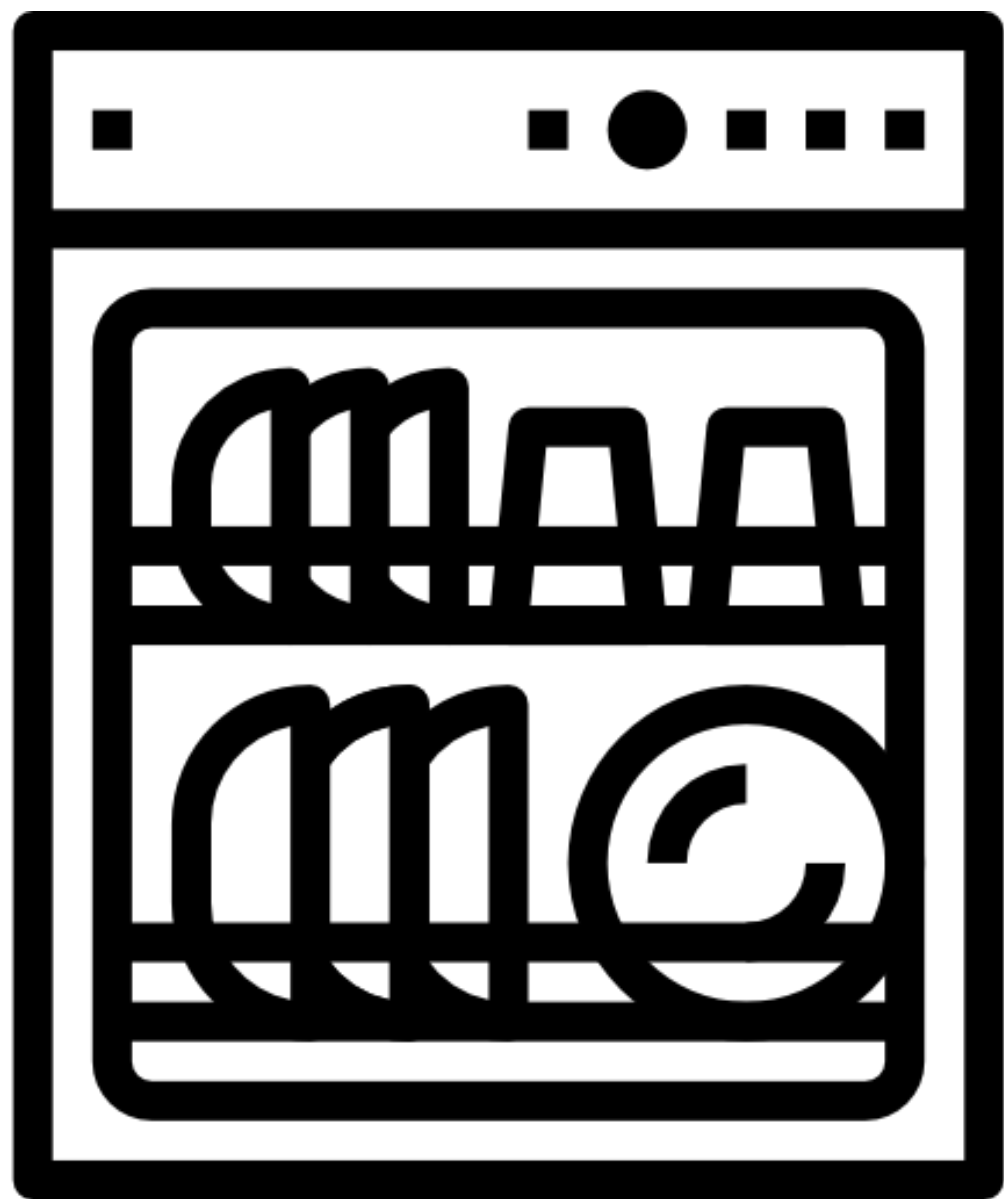
Efficient & Predictable AI via Mathematical Structures in Data



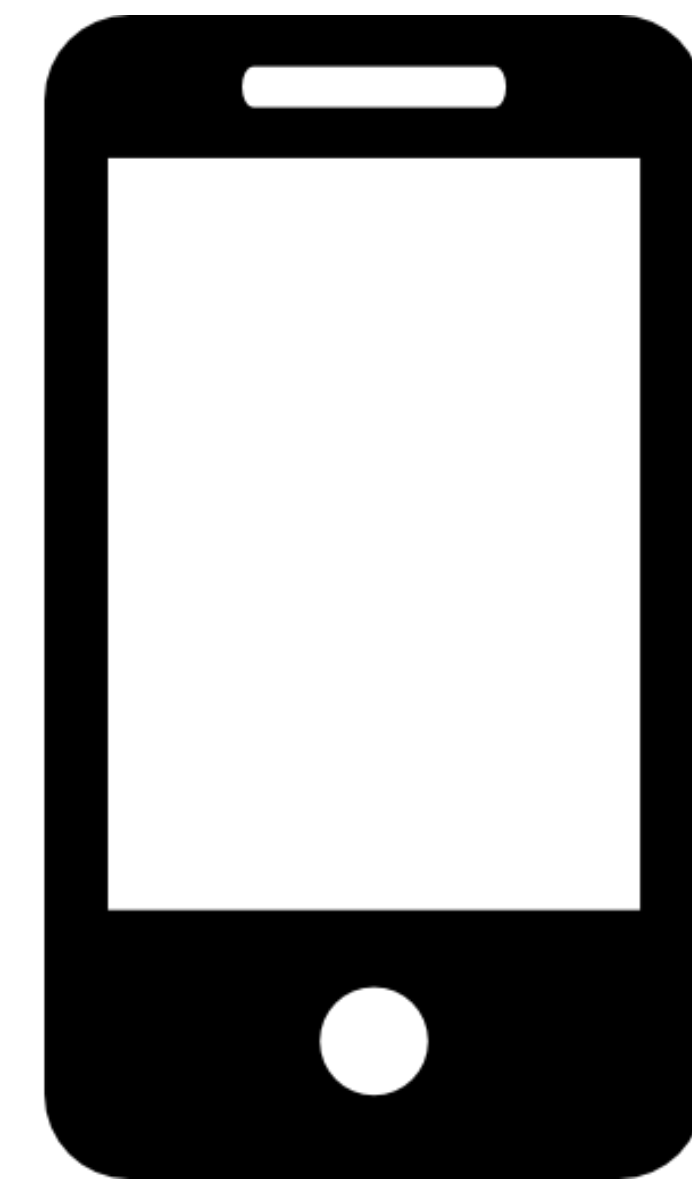
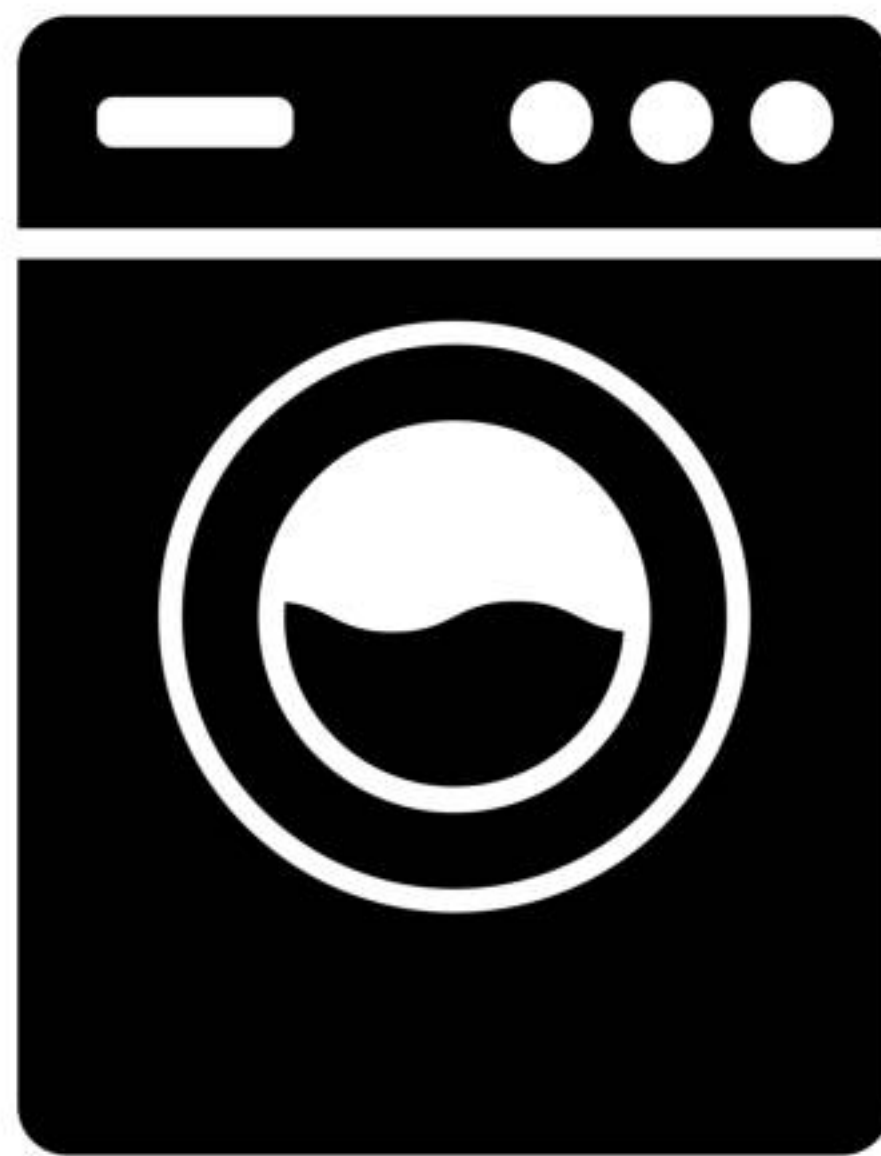
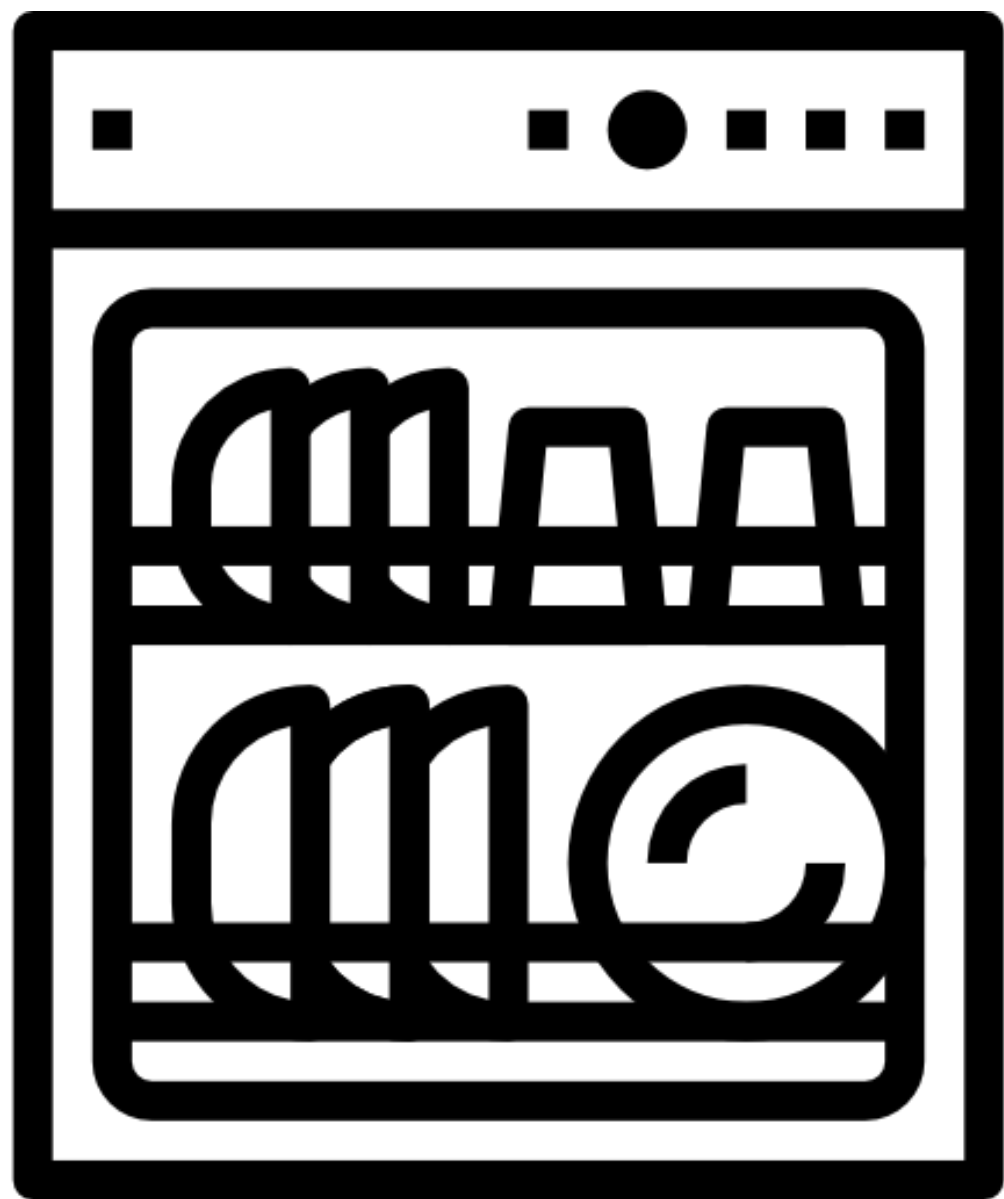
work with Victoria Kauffman, Aadit Saluja, Florian Klein, Jeremy Hsu, David Alvarez-Melis, Despina Argiropoulos, Cathy Hou, Vlad Cainamisir, Qitong Wang, Konstantinos Kopsinis, Stratos Idreos

From Genaral Intelligence to Specialized Intelligence

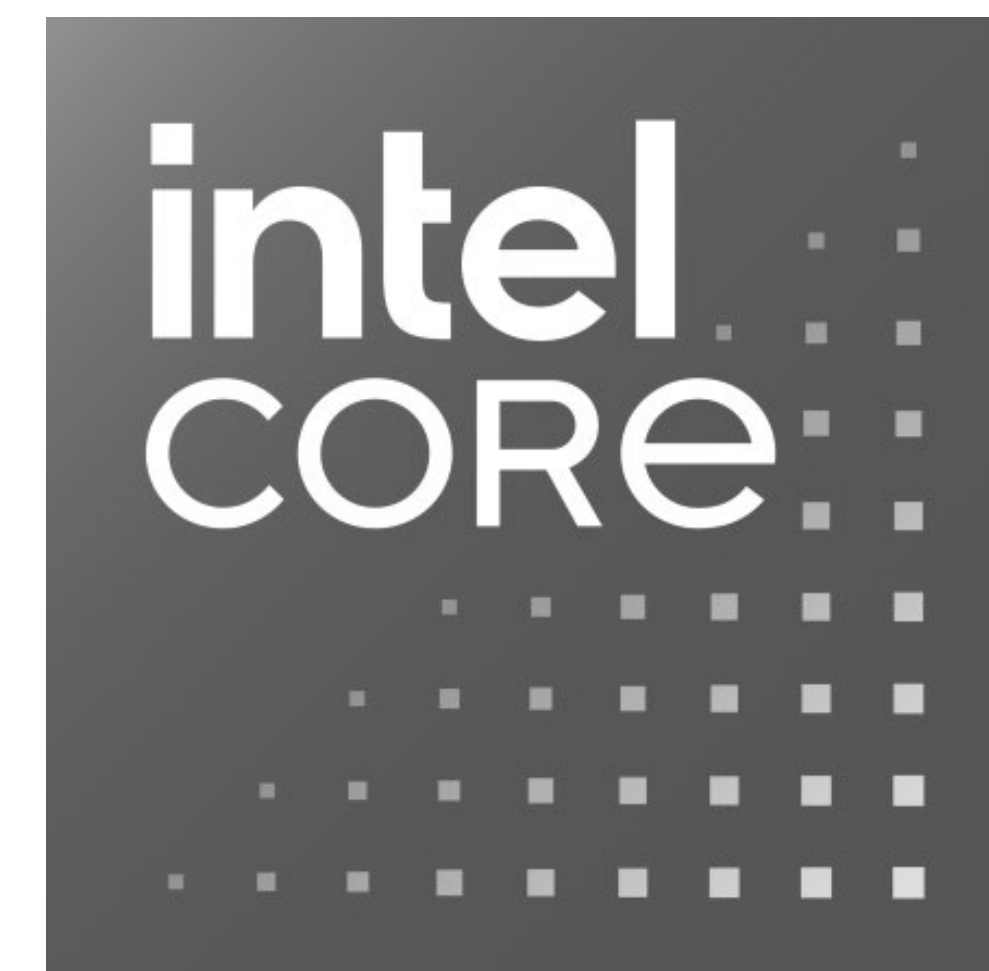
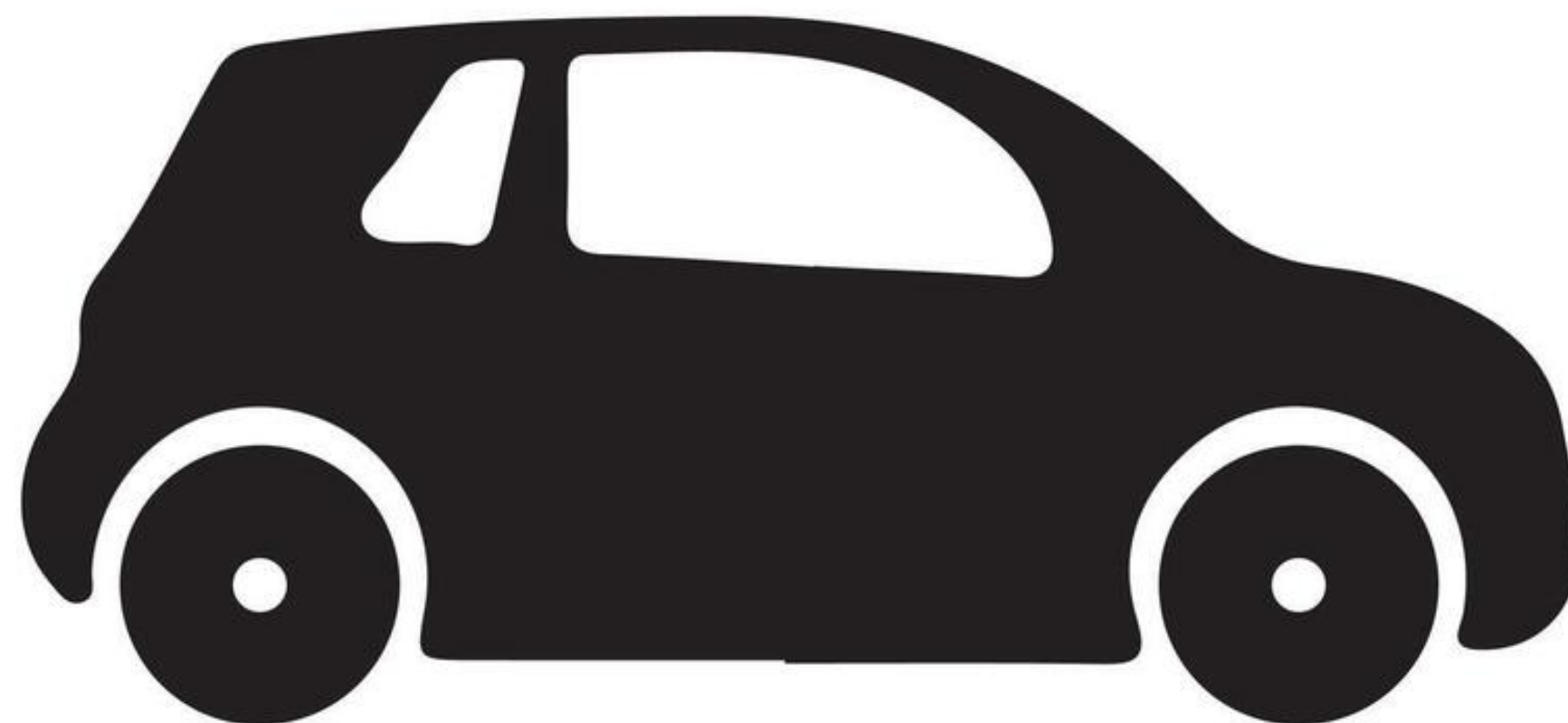
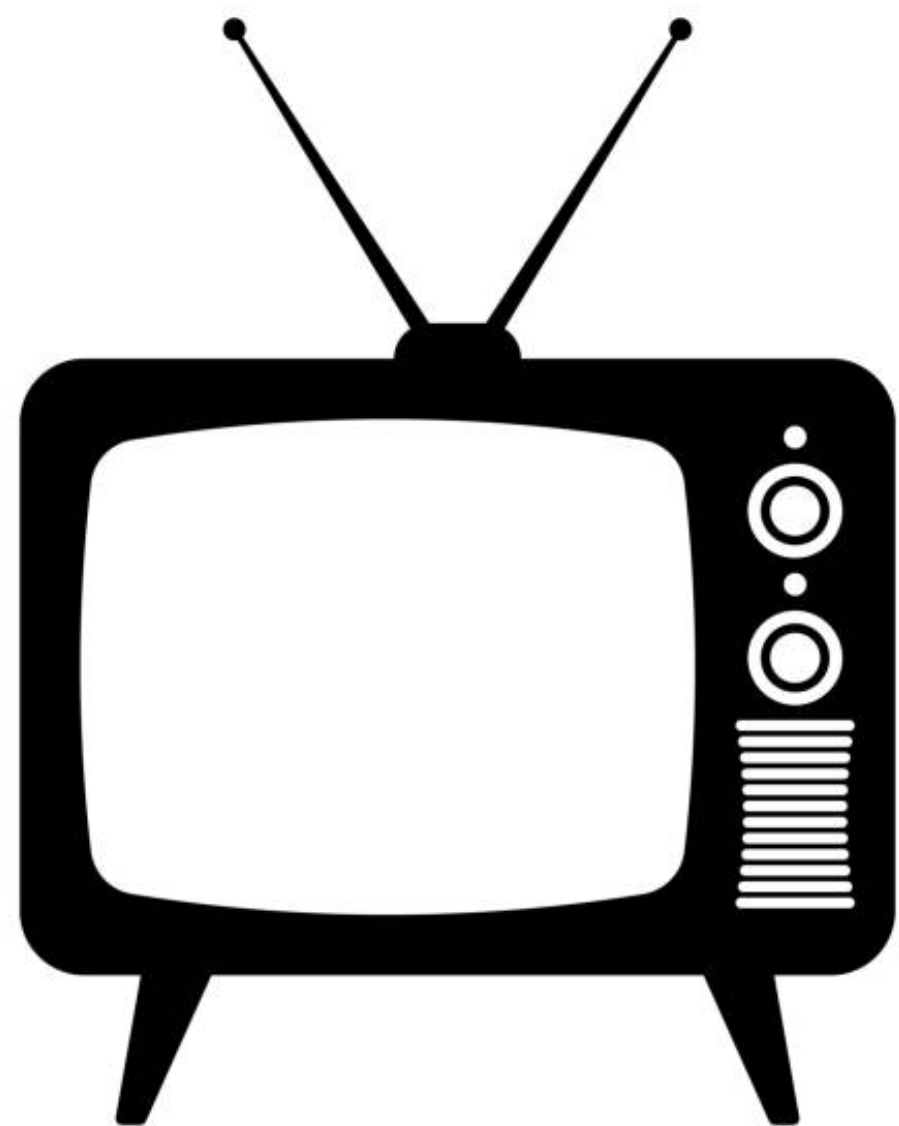
We don't know how they work

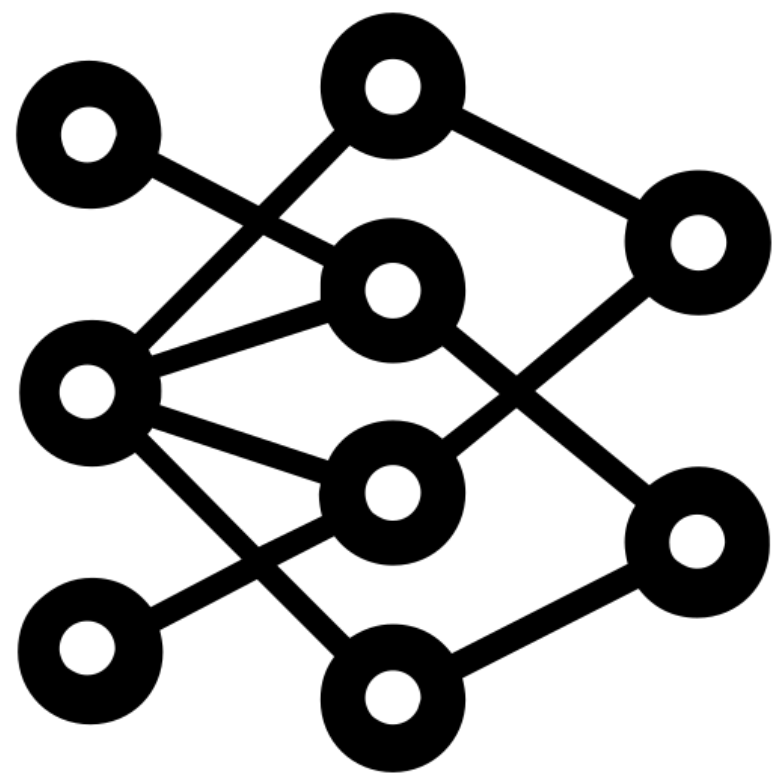


We don't know how they work

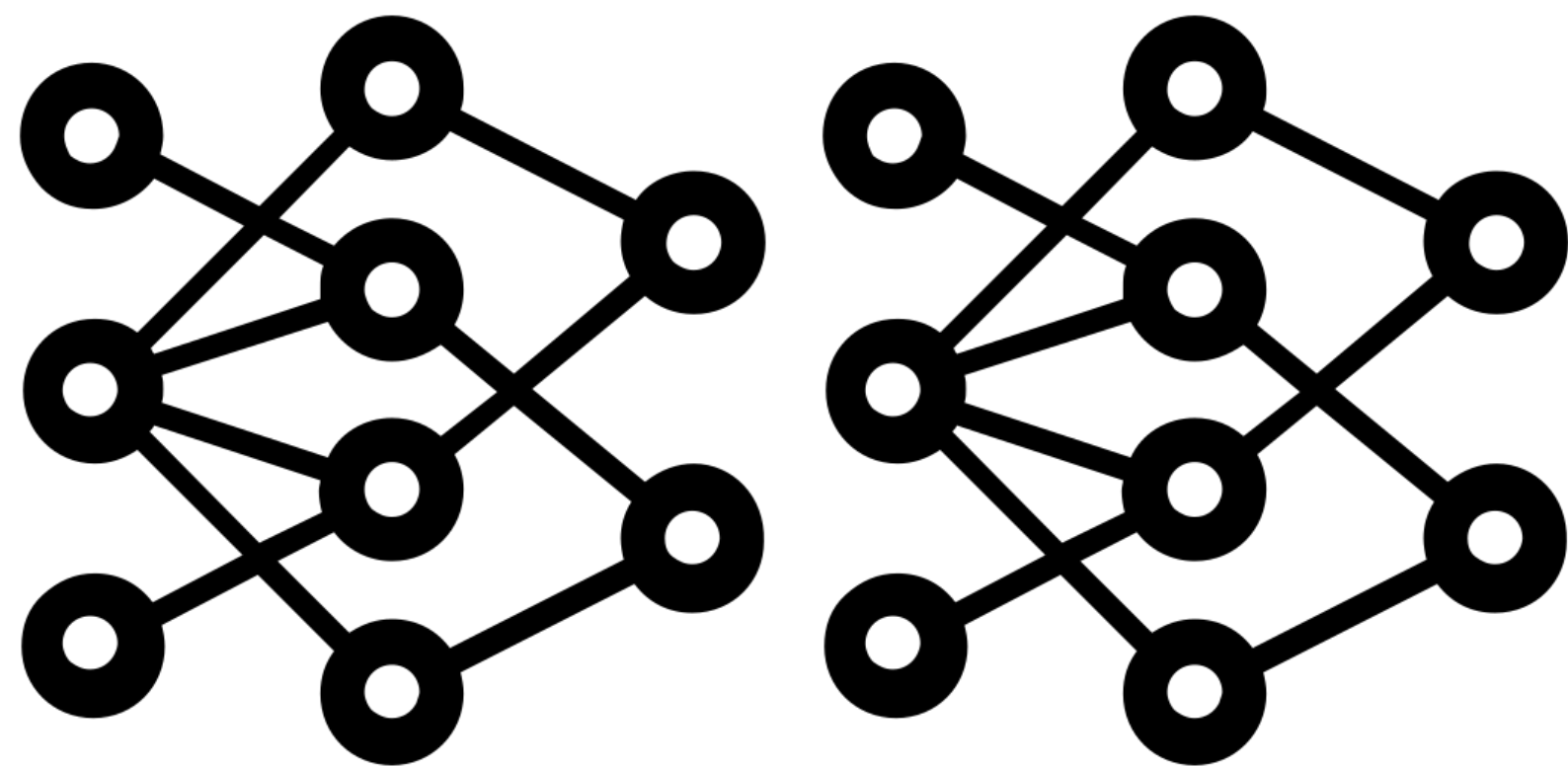


We don't know how they work

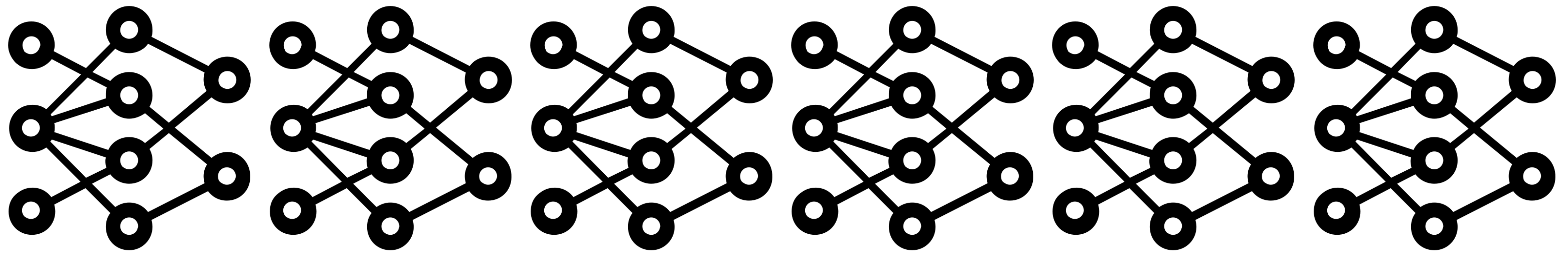




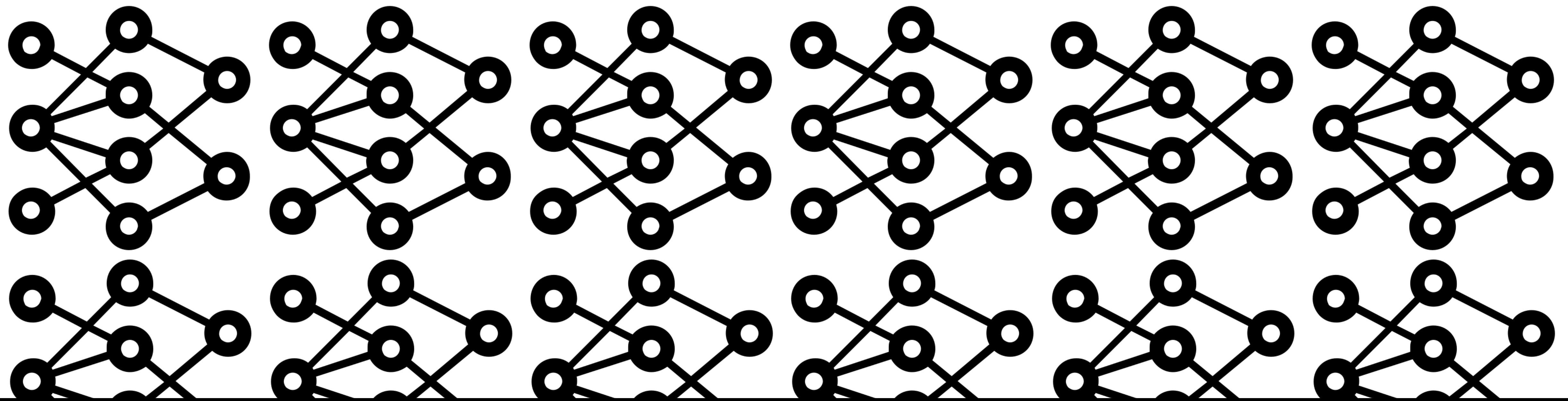
We don't know how they work



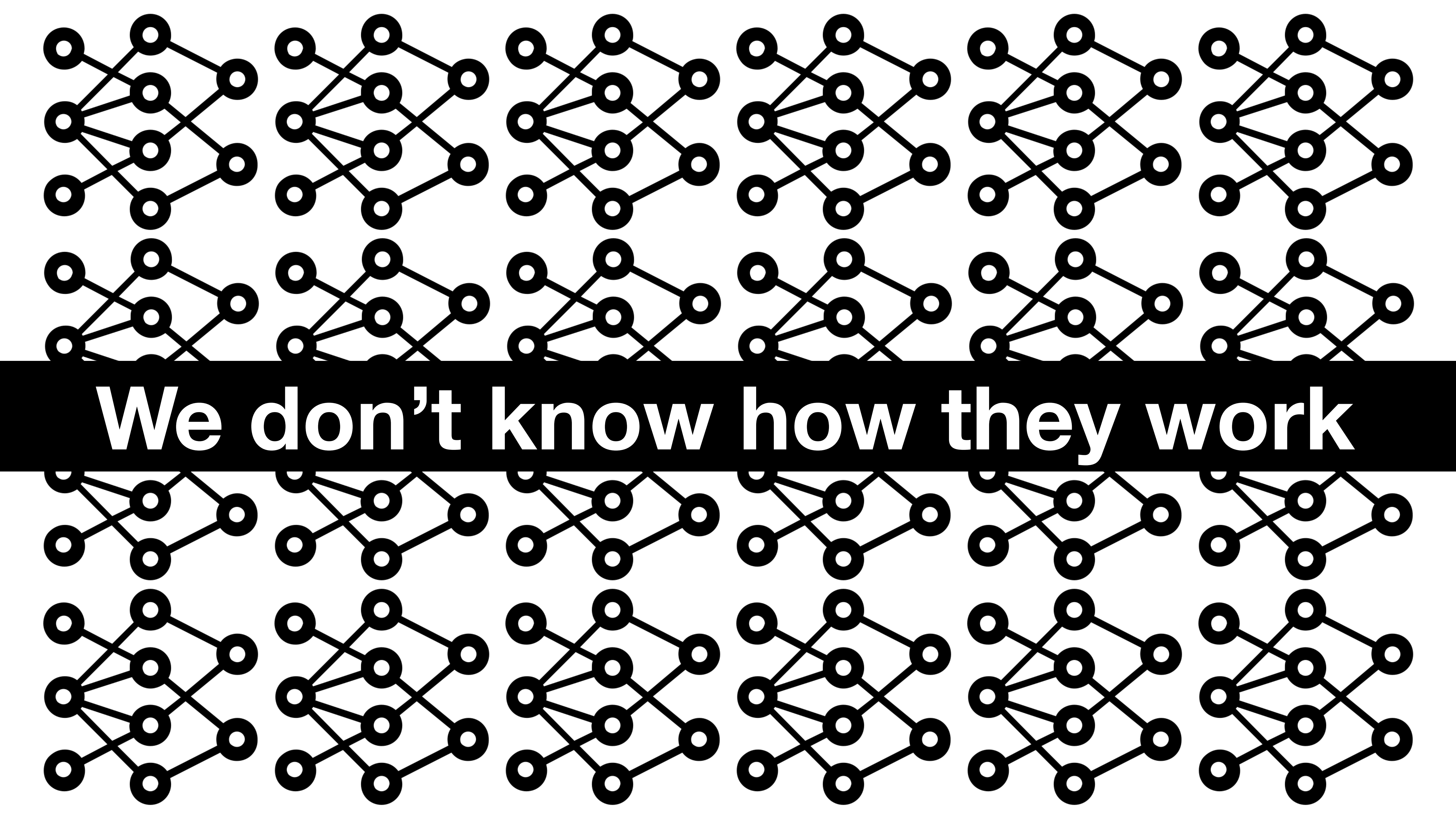
We don't know how they work



We don't know how they work



We don't know how they work



We don't know how they work

Artificial *General* Intelligence

What's on your mind today?

+ Ask ChatGPT

✻ Hey, early bird

How can I help you today?

Sonnet 5.5 Medium



Your move, Utku!

+ Ask Gemini

How can I help you?

+ Ask Qwen

Auto ▾



Slow, inaccurate, unreliable, unpredictable...

how many d's are there in Wednesday?

There is **1** "d" in **Wednesday**.



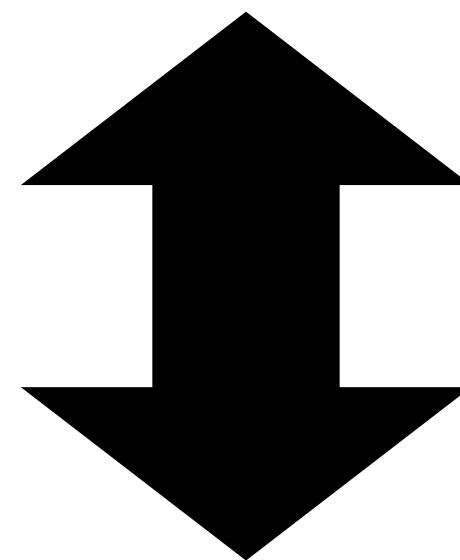
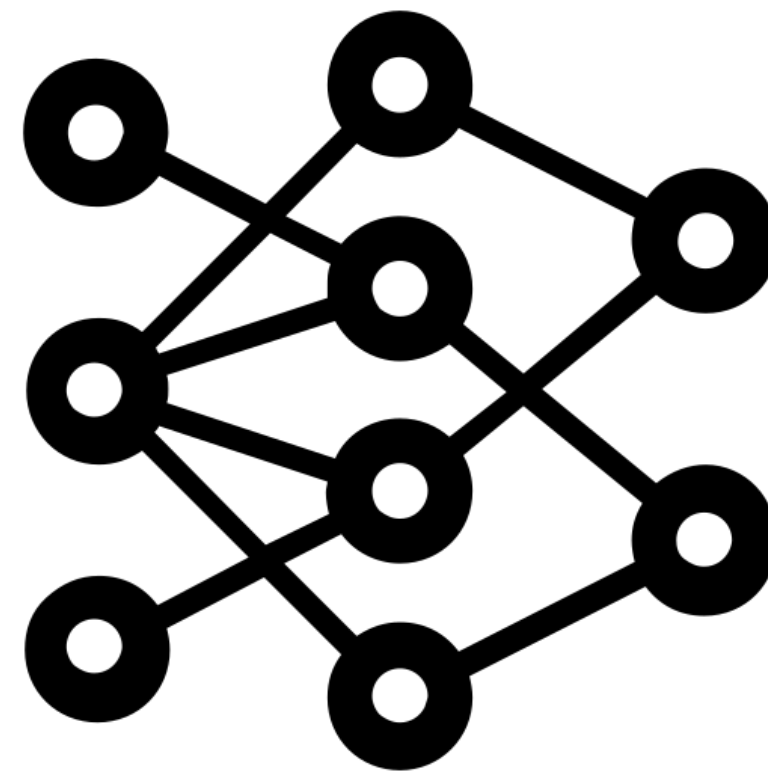
Co-Design

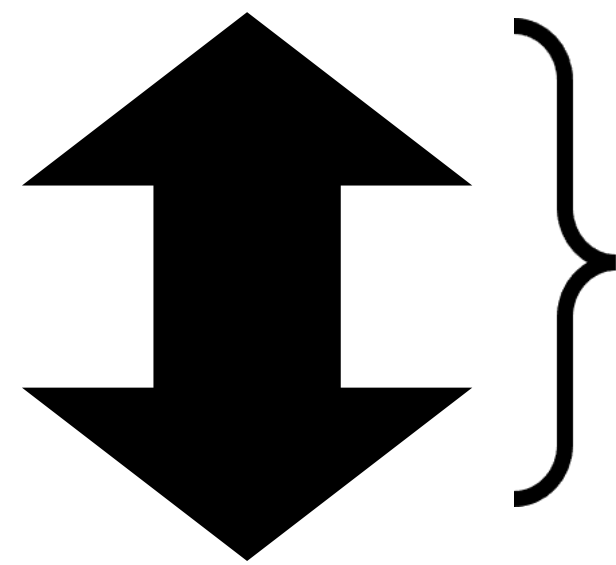
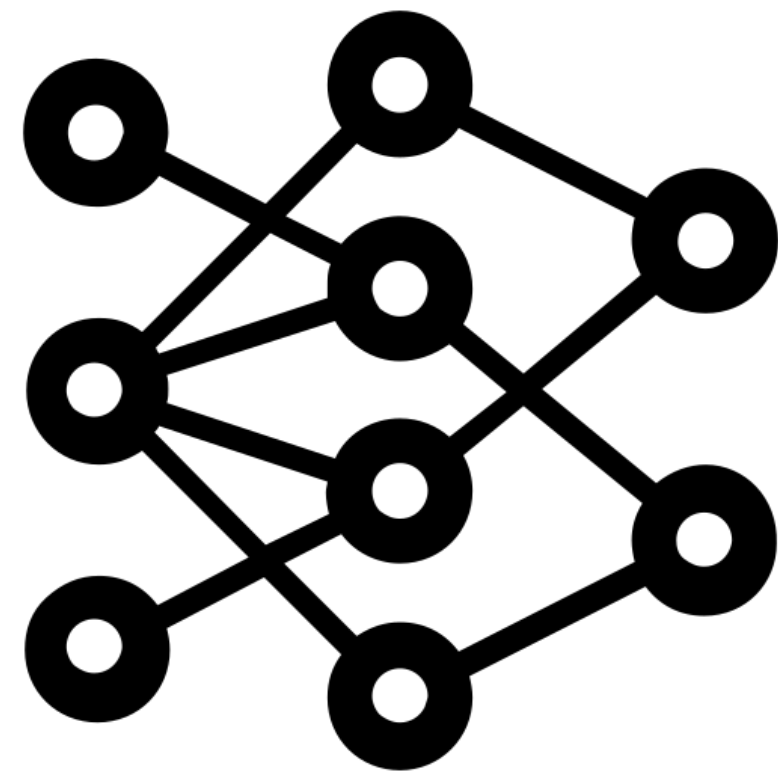
=

Specialization

=

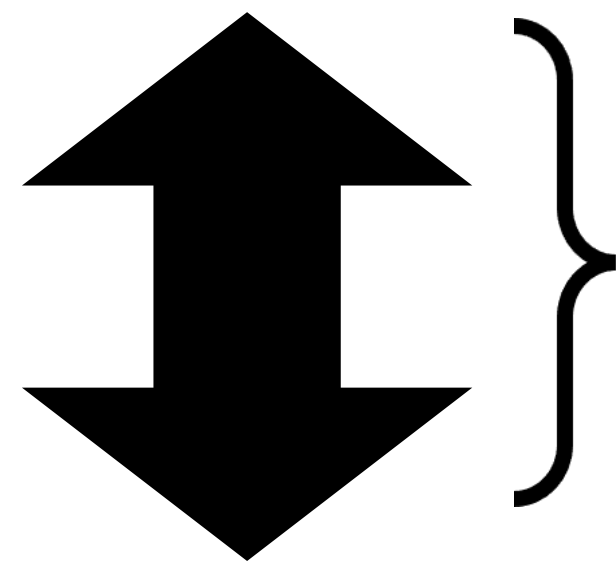
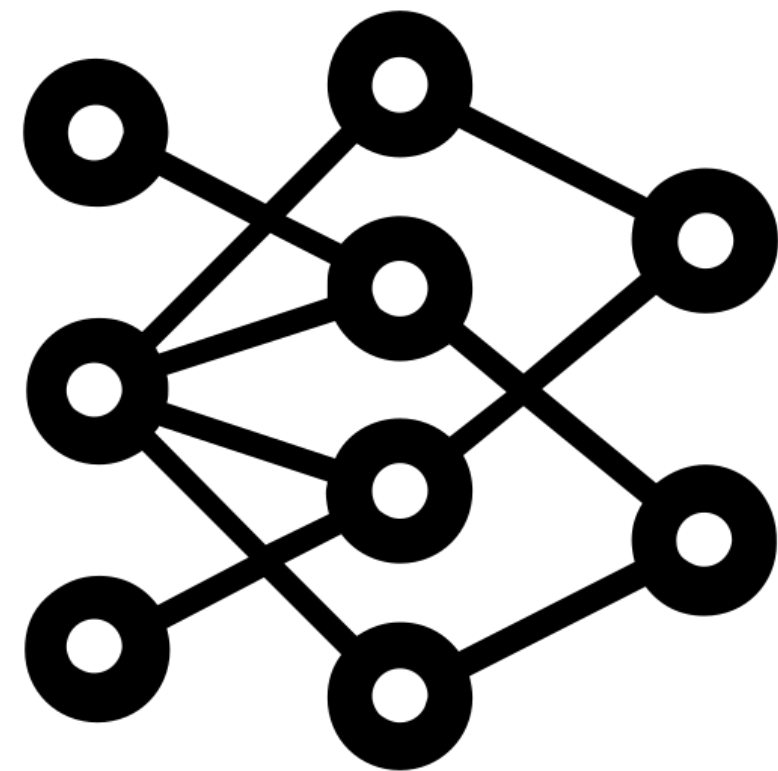
Vertical Integration





Reasoning Mechanisms

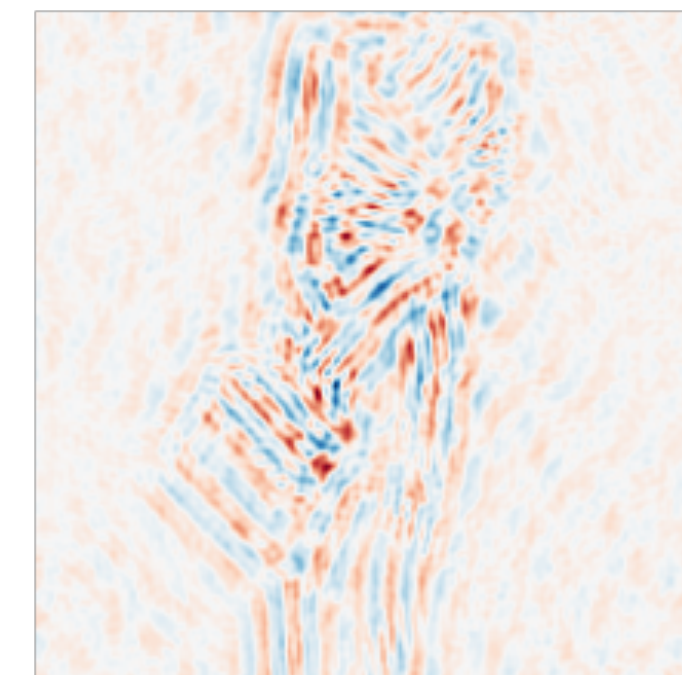
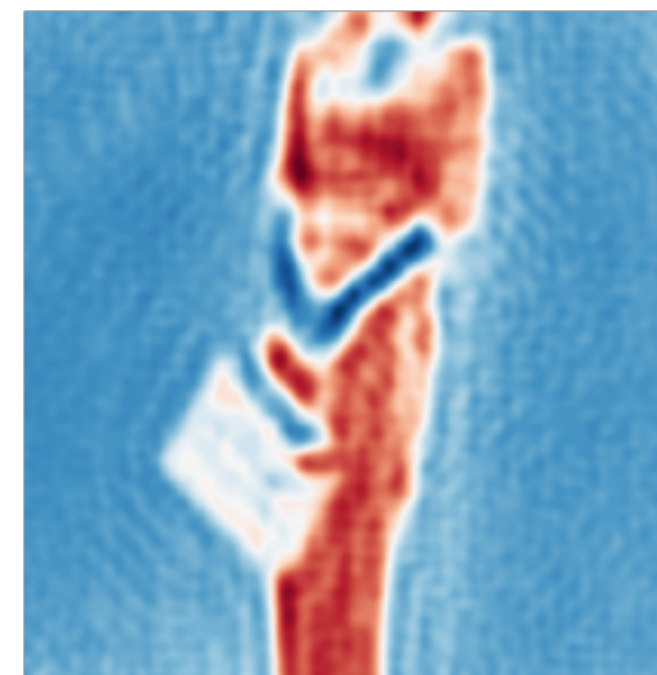


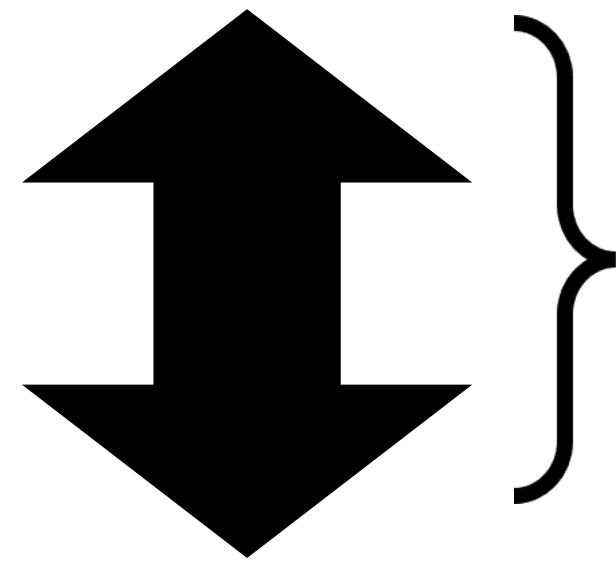
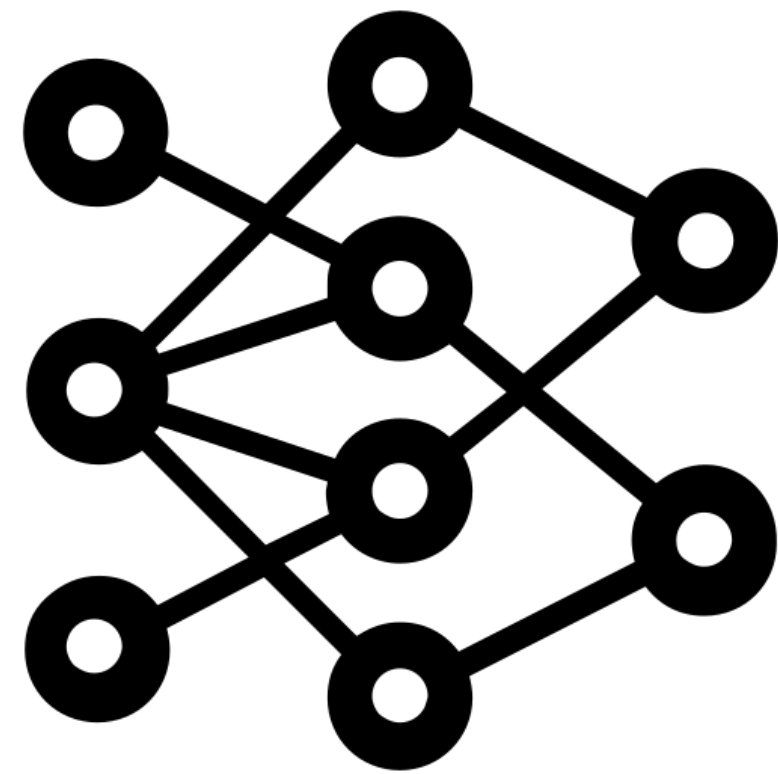


***Reasoning
Mechanisms***



***Workload
Characterization***

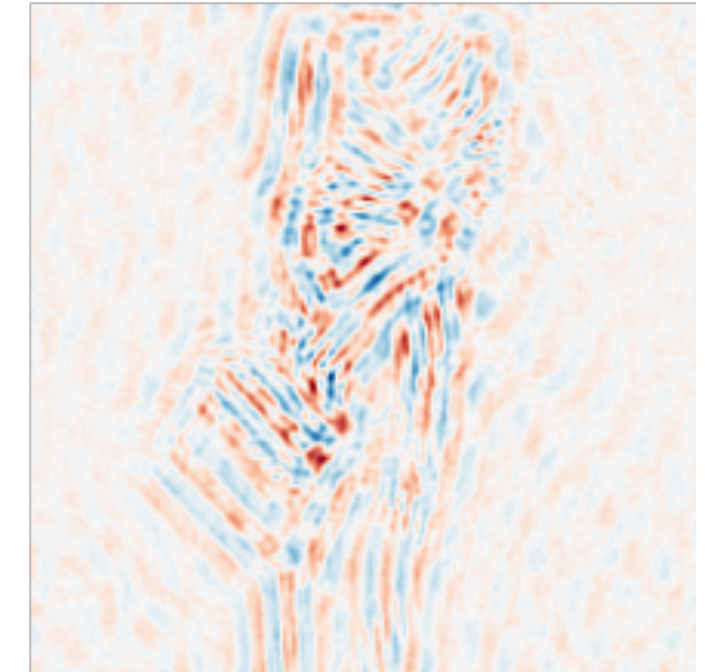




*Reasoning
Mechanisms*



*Workload
Characterization*



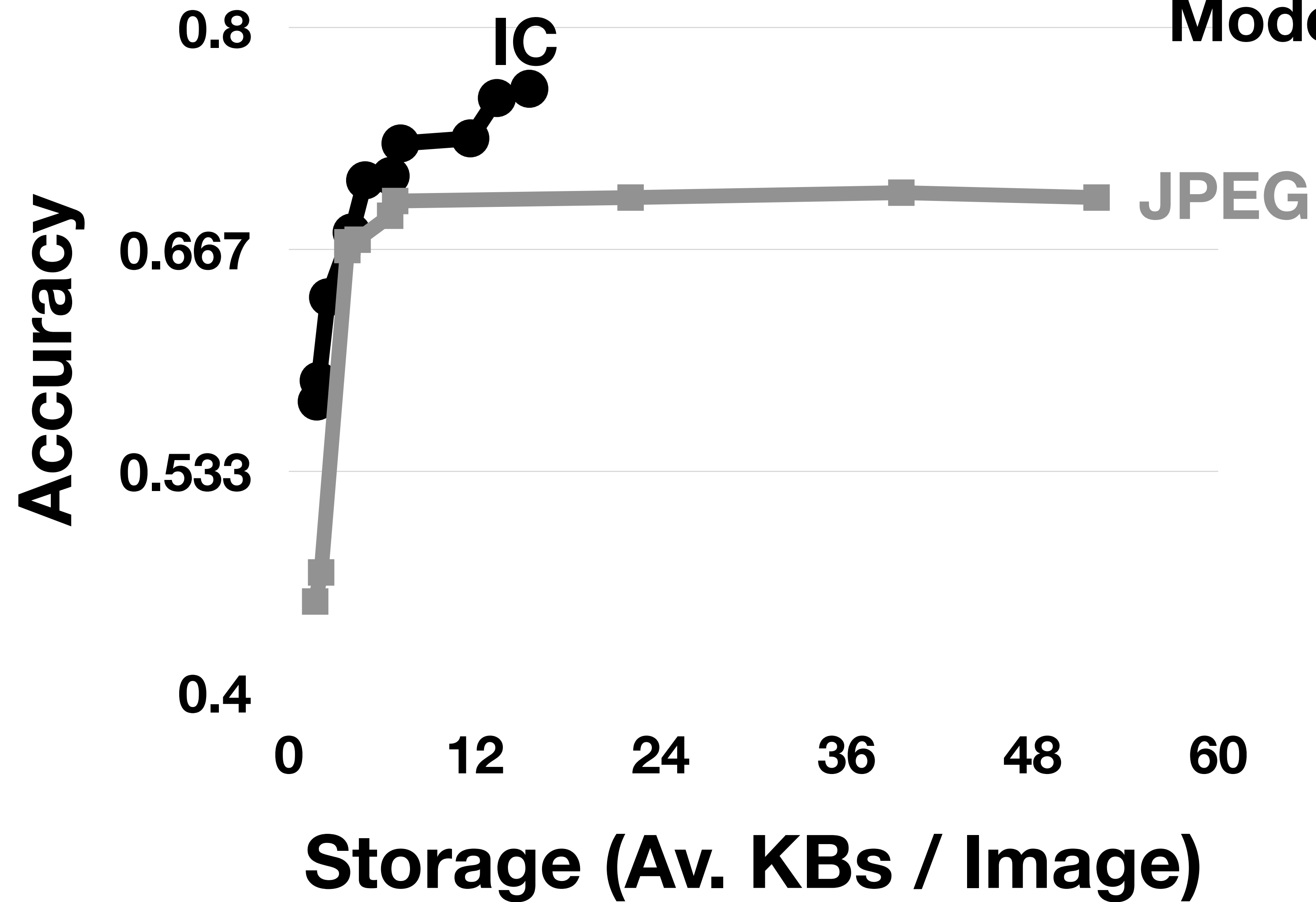
Data Engineering



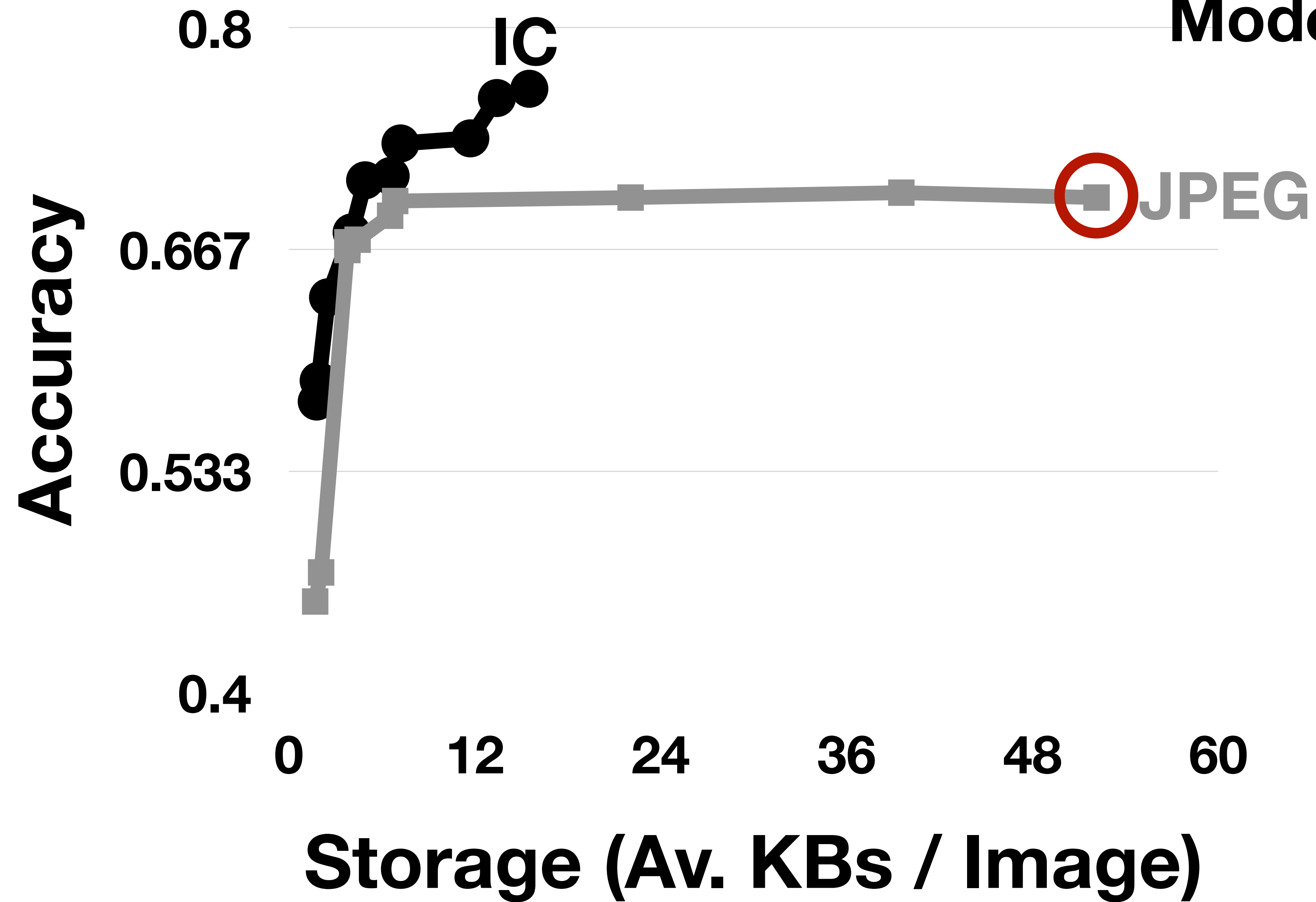
		Definition	Domain	A new design
Design Primitives	Pruning	It defines a mask that sets values of data items to zero.	$k \in \{0, \dots, 1/8 \times B_P\}$ $M_k(i, j) = \begin{cases} 0, & i + j < k, \\ 1, & i + j \geq k \end{cases}$	M_2 (prune two top-left diagonals)
	Quantization	It defines a map where every value in the image is divided by a constant and rounded to its nearest integer.	$\mathbf{q} = (q_l, q_m, q_h) \in \{1, 10, 40, 100\}^3$ $Q_q(i, j) = \begin{cases} q_l, & i + j < B_Q/4 \\ q_m, & B_Q/4 \leq i + j < B_Q \\ q_h, & B_Q \leq i + j \leq 2B_Q - 2 \end{cases}$	$Q_{40,10,100}$ (quantize low-, mid-, high-frequency zones by 40, 10, and 100)
	Downsampling	It defines reducing spatial dimensions of an image.	$DF \in \{1, 1.2, 1.4, 1.6, 1.8, 2, 2.25, 2.5, 2.75, 3, 3.5, 4\}$	1.4 (downsample the image by 1.4×1.4)
	Upsampling	It defines increasing spatial dimensions of an image.	$UF \in \{1, 1.2, 1.4, 1.6, 1.8, 2, 2.25, 2.5, 2.75, 3, 3.5, 4\}$	2.2 (upsample the image by 2.2×2.2)
	Block Size	It defines spatial granularity of pixels to store/represent images.	$B = (B_{DS} \times B_{DS}, B_{P,Q,US} \times B_{P,Q,US})$ $B_{DS} = 224, B_{P,Q,US} = \lfloor 224/DF \rfloor$	(224x224, 160x160) (160 is due to DF of 1.4)
	Color	It defines which channels compression primitives should be applied to.	$C \in \{(C_Q, C_P)\}$ $C_Q \in \{YCrCb, CrCb\}, C_P = YCrCb$	$\{CrCb, YCrCb\}$ (quantize only color channels)
	Order	It defines in which order compression primitives should be applied to.	$O = \text{downsample} \rightarrow \text{prune} \rightarrow \text{quantize} \rightarrow \text{upsample}$	downsample $\rightarrow$ prune $\rightarrow$ quantize $\rightarrow$ upsample
	Encoding	It defines lossless encoding compression level of data.	$CL = 7$	7 (use compression level of 7)

Dataset: GTOS

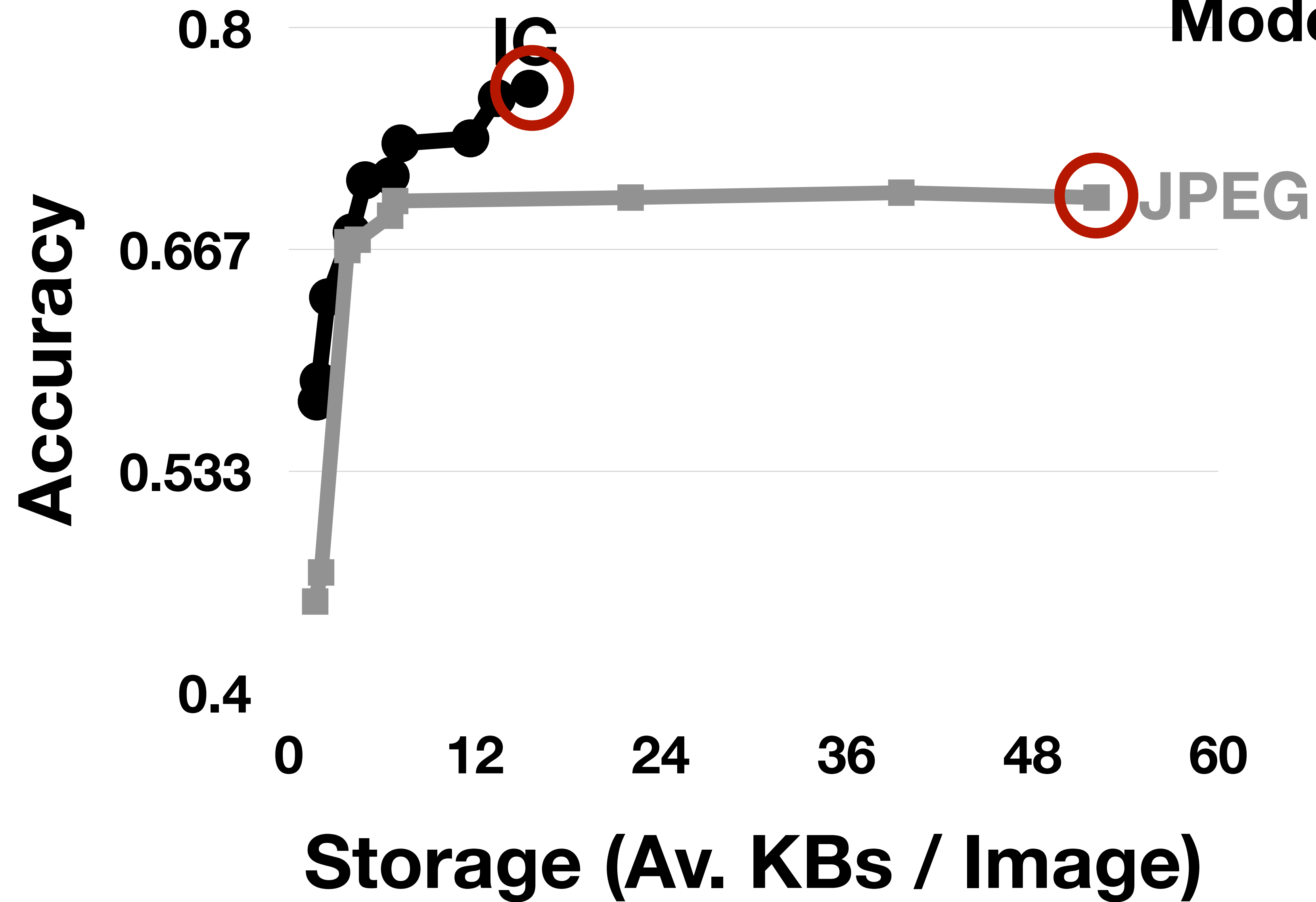
Model: ResNet-50



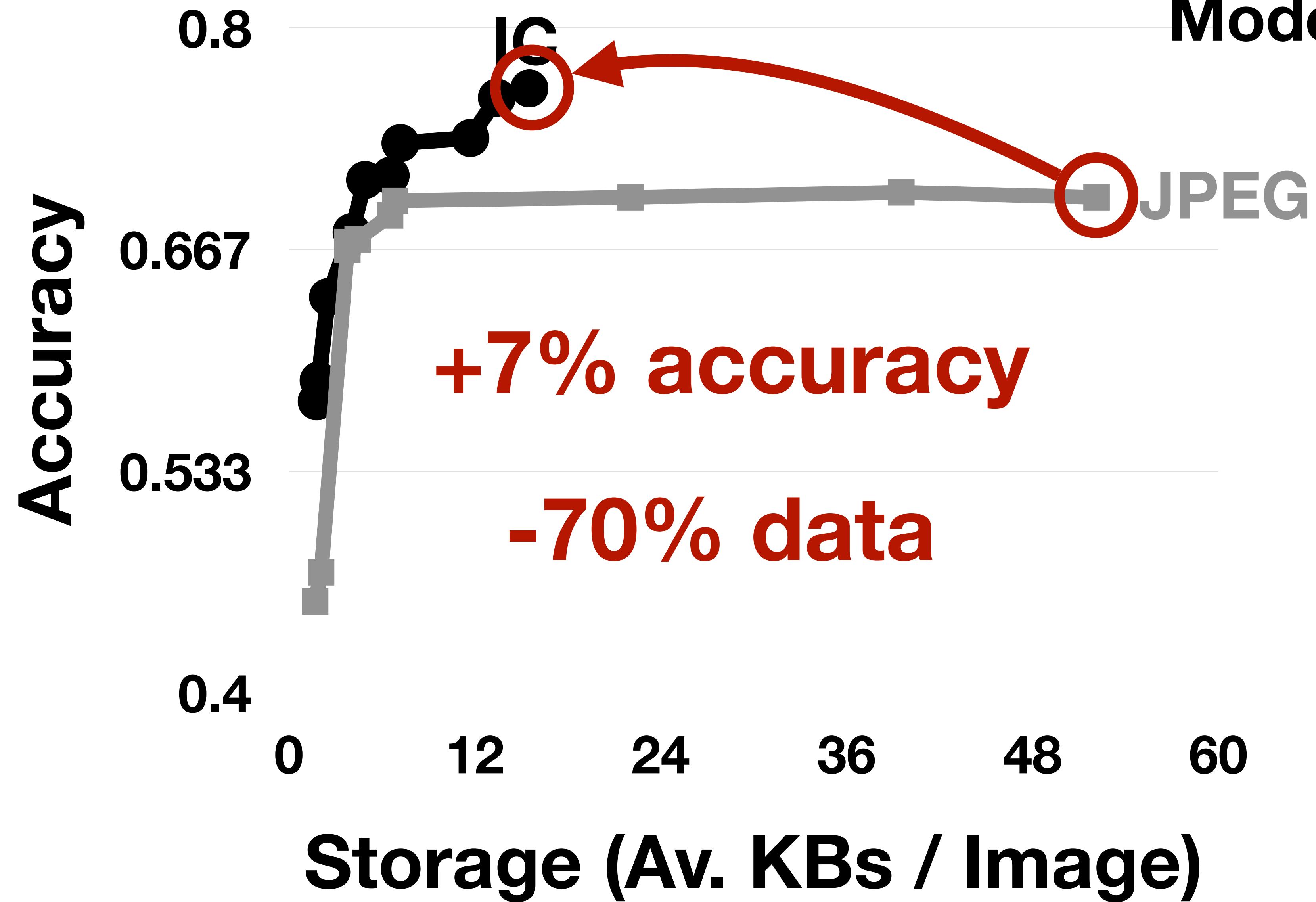
Dataset: GTOS
Model: ResNet-50



Dataset: GTOS
Model: ResNet-50



Dataset: GTOS
Model: ResNet-50



how many d's are there in Wednesday?

There are **2 d's** in **Wednesday**.



Are you sure?

Yes — **Wednesday** has **2 d's**:

W e d n e s d a y

The **d** appears twice.



Are you really sure?

Yes. **Wednesday** contains **2 letter d's**:

Wednesday.



Thanks!