BUILDING (AND BREAKING) MACHINES THAT THINK FAST AND SLOW

Tom Goldstein

UNIVERSITY OF MARYLAND

OVERVIEW

What are adversarial attacks?

What can adversarial attacks do for you?

Can neural nets “think”?

ADVERSARIAL ATTACKS

"Egyptian Cat" 28%

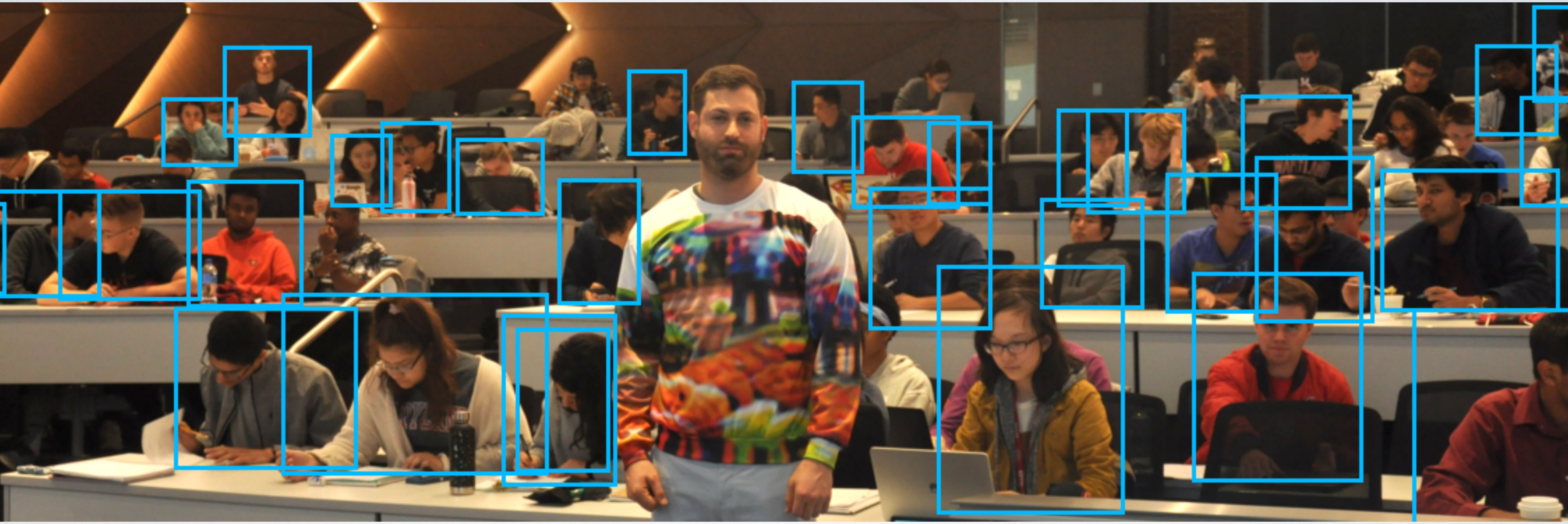
"Traffic Light" 97%

How far can these attacks go?

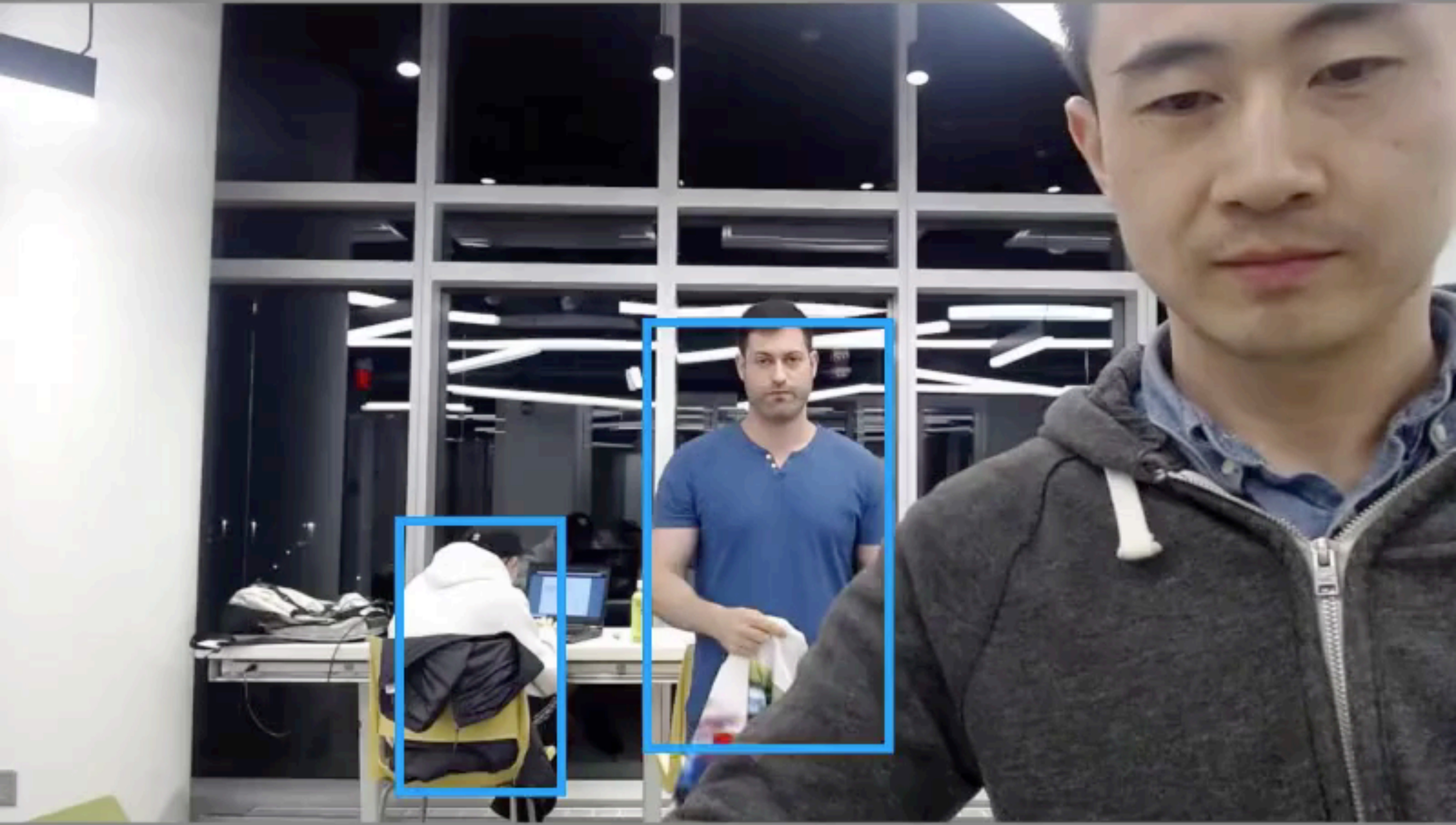
vs

ADVERSARIAL ATTACKS

Yolov2 Object Detections



Wu, Lim, Davis, G. "Building an invisibility cloak"



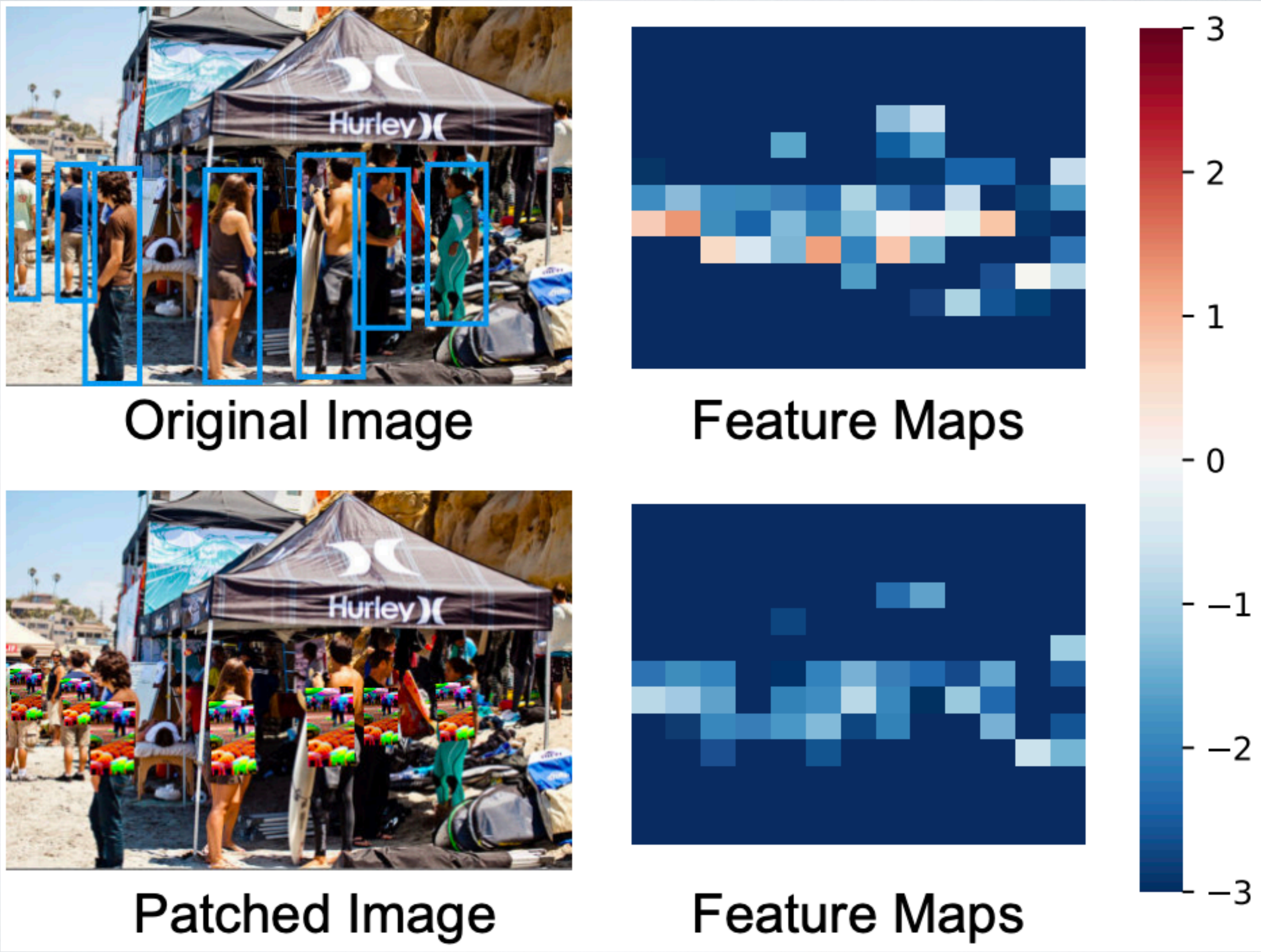
THE NEW YORKER

"[The Cloak] looks like a baggy sweatshirt…
with garish colors in formless shapes."

THE TIMES

"This hideous jumper makes Professor Goldstein invisible…
…to the fashion curators at Vogue."

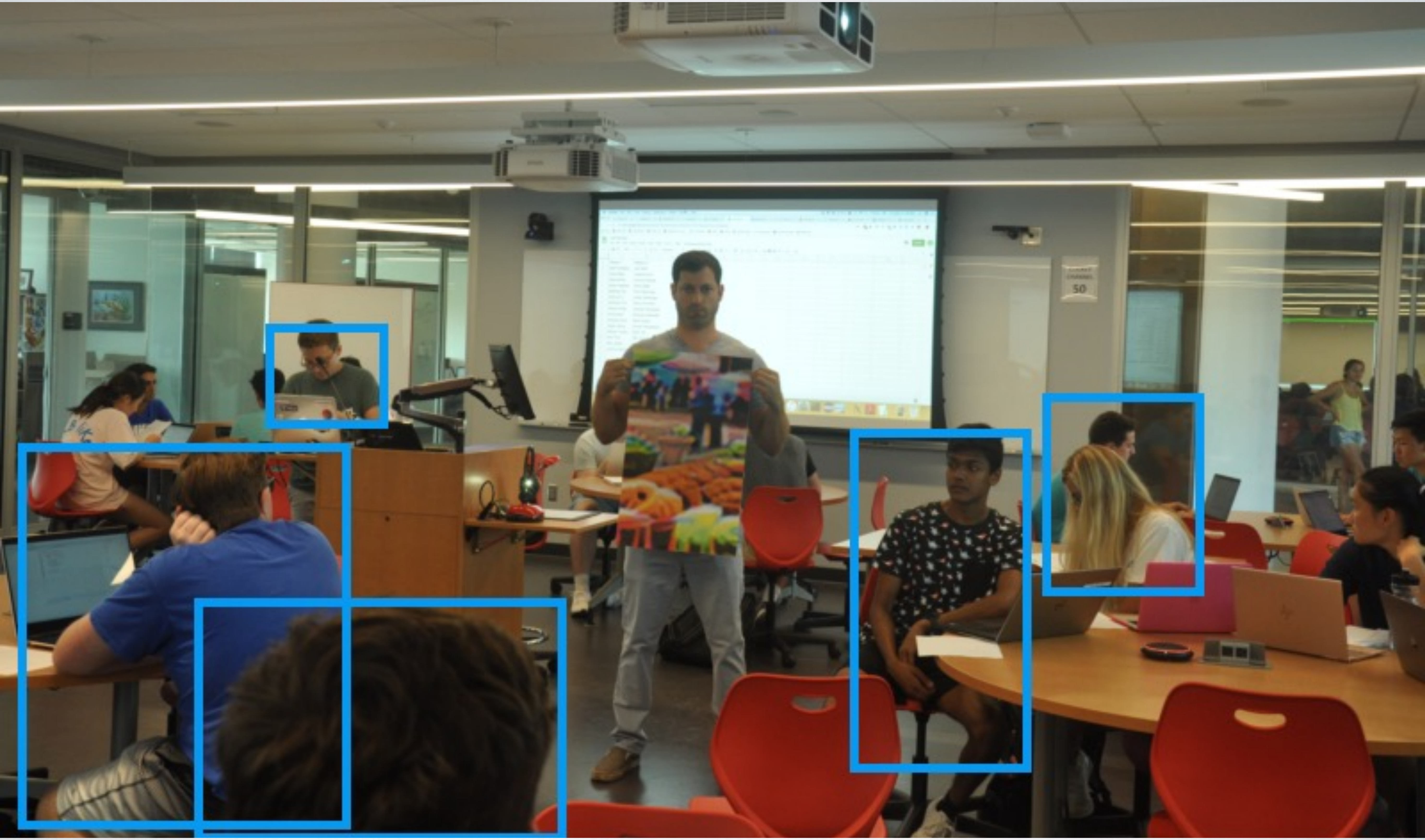
APPROACH



Hurley
Original Image
Feature Maps
Patched Image
Feature Maps
3
2
1
0
−1
−2
−3

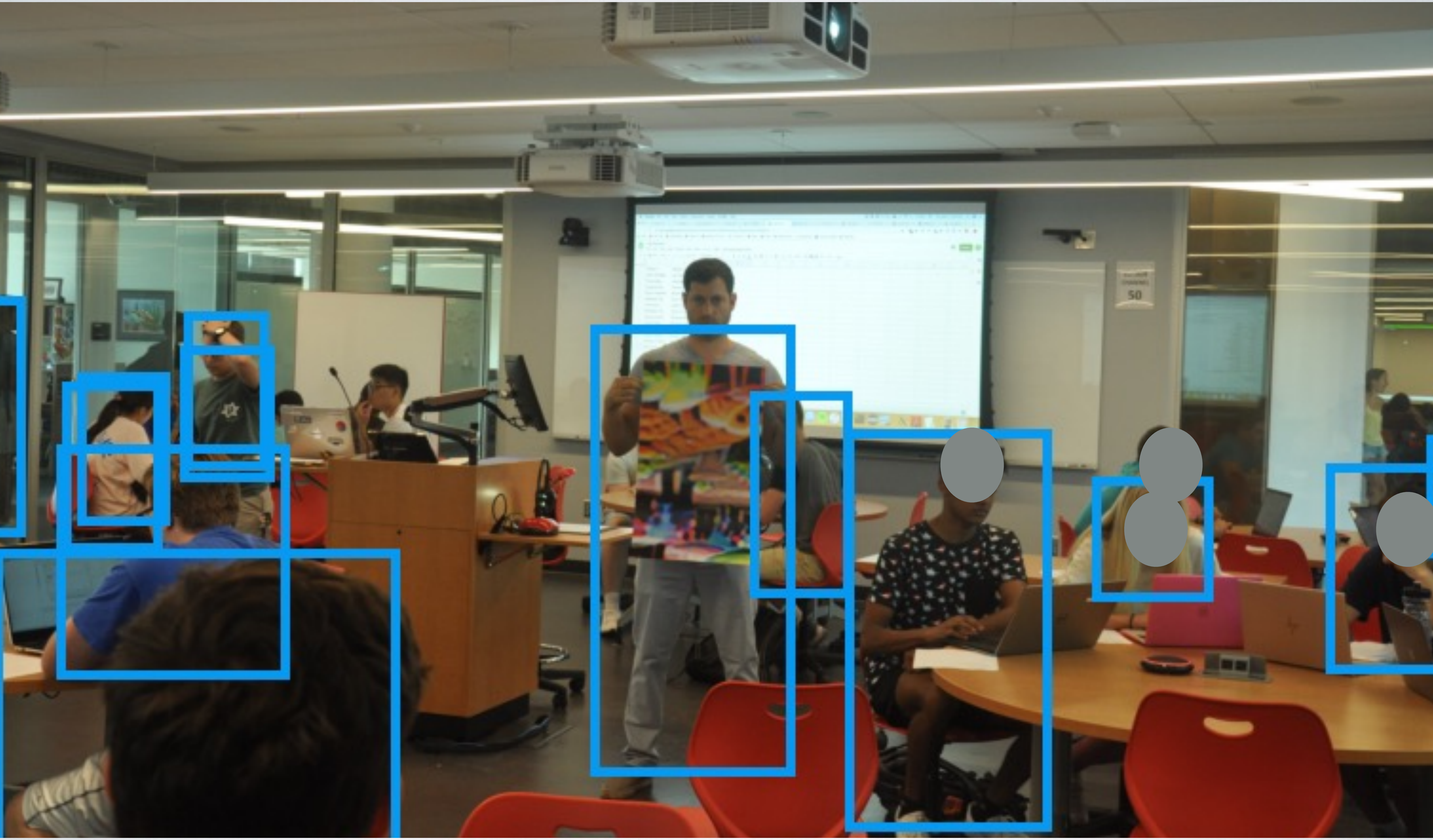
THE SWEATER TEST

THE FLIP TEST



Wu, Lim, Davis, G. "Building an invisibility cloak"

THE FLIP TEST



Wu, Lim, Davis, G. "Building an invisibility cloak"

Other work on breaking systems

Adversarial attacks on copyright systems

Saadatpanah, Shafahi, & Goldstein

Witches' Brew: Industrial Scale Data Poisoning via Gradient Matching

Geiping, Fowl, Huang, Czaja, Taylor, Moeller, Goldstein

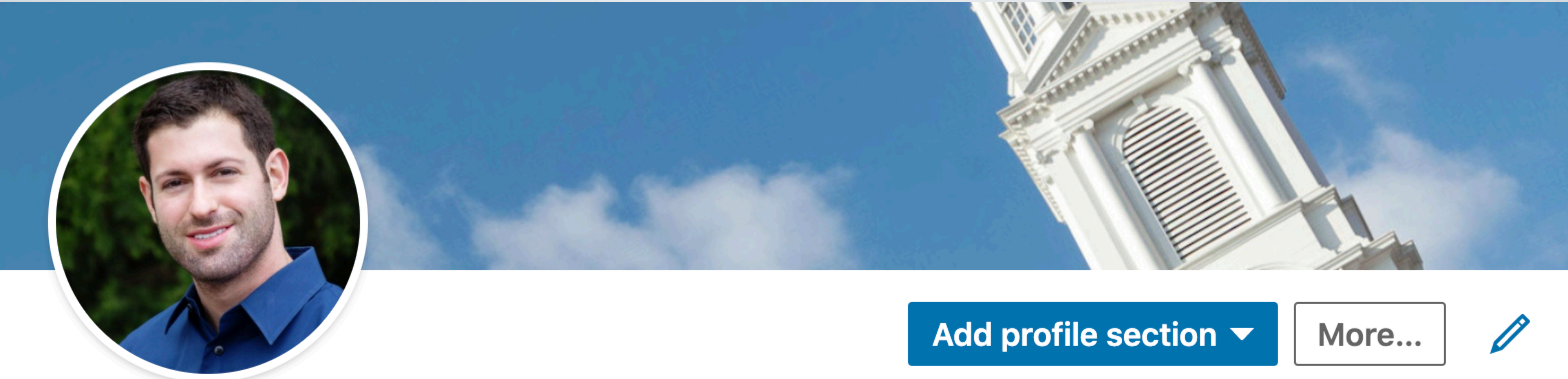
Cloud AutoML Vision

Adversarial Attacks on Machine Learning Systems for High-Frequency Trading

Goldblum, Schwarzschild, Patel, Goldstein

Can adversarial ML protect **privacy**?

Linked in



Add profile section ▾ | More...

Tom Goldstein

Associate Professor at University of Maryland

Washington, District Of Columbia · 263 connections ·

Contact info

University of Maryland

University of California, Los Angeles

About

Tom is an expert on large-scale and distributed optimization methods for machine learning, computer vision, and signal processing. Areas of focus include:

• Machine learning and AI... see more

LinkedIn

Clearview.AI

Private companies

Political organizations

Law enforcement

Foreign governments

Can we poison datasets so that they’re useless?

Can we poison datasets so that they're useless?

Related work

Huang, *Unlearnable Examples*, 2021

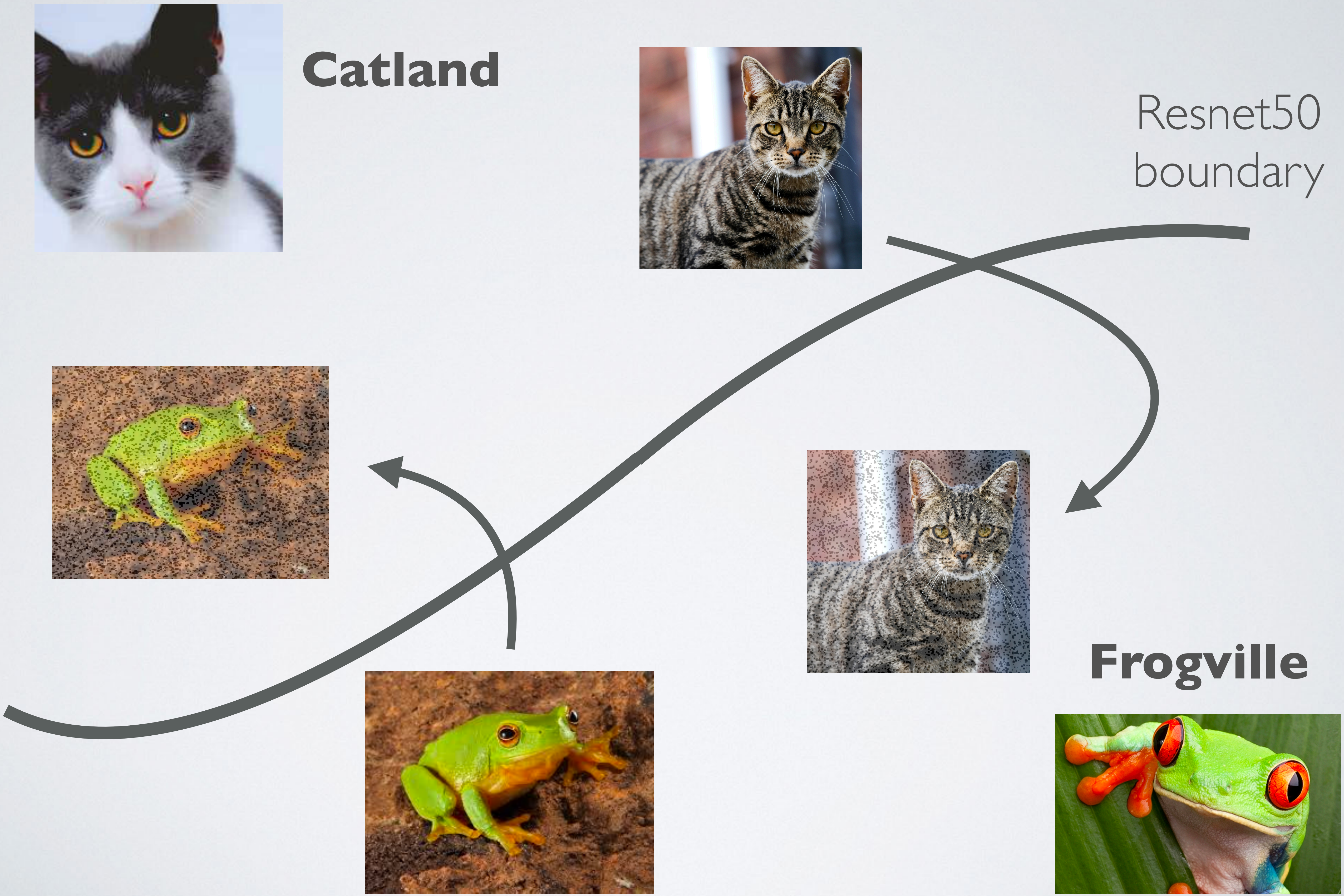
Shen, *TensorClog*, 2021

Fowl & G, *Preventing Unauthorized use*, 2021

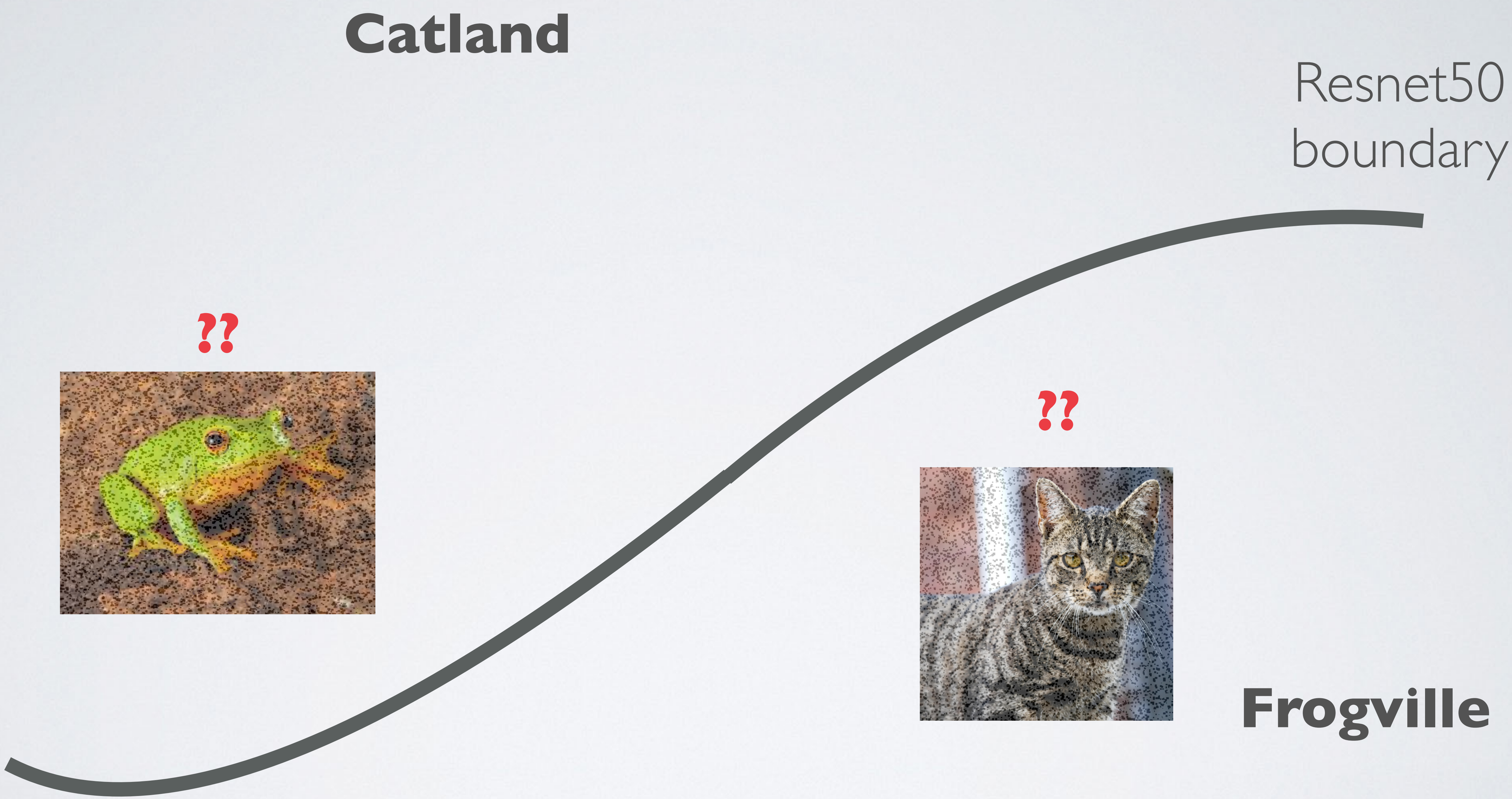
Yu, *Indiscriminate Poisoning*, 2022

Sandoval-Segura & G, *Autoregressive Perturbations*, 2022

TRAINING ON ADVERSARIAL EXAMPLES



OVER/UNDER PARAMETERIZED DUALITY



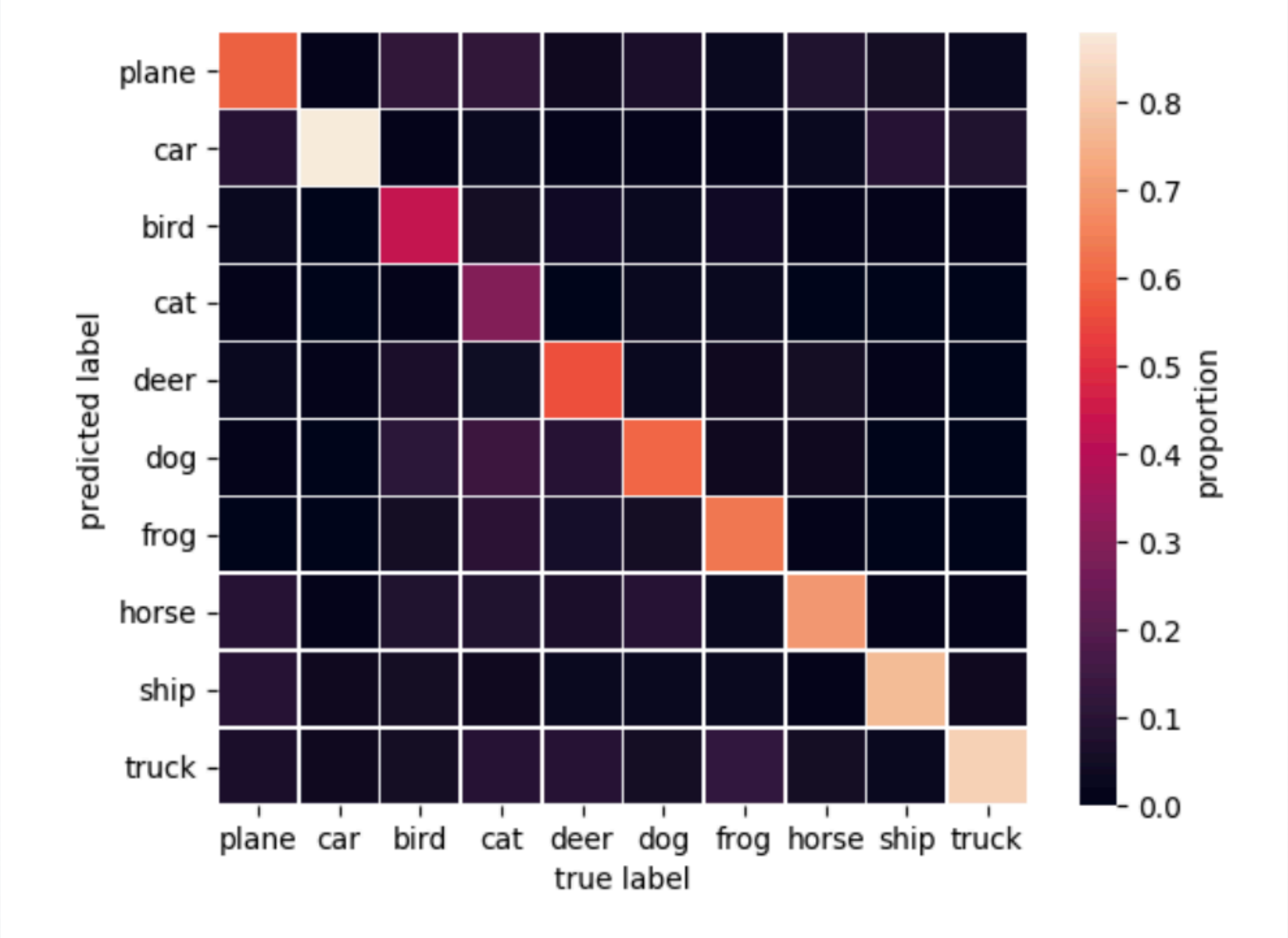
OVER/UNDER PARAMETERIZED DUALITY

TRAIN ON ADVERSARIAL EXAMPLE
TEST ON CLEAN DATA

Base image

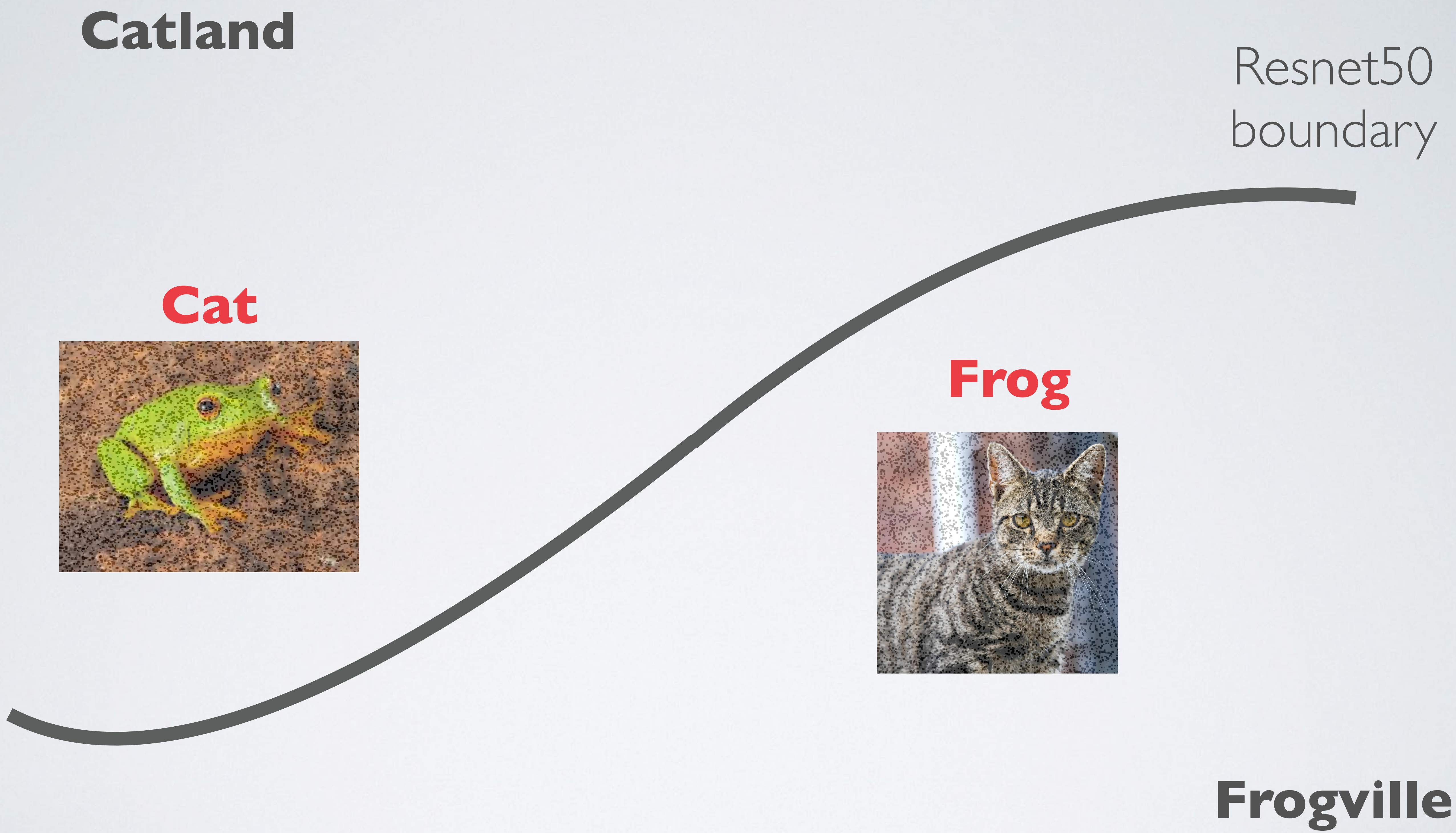
Cat

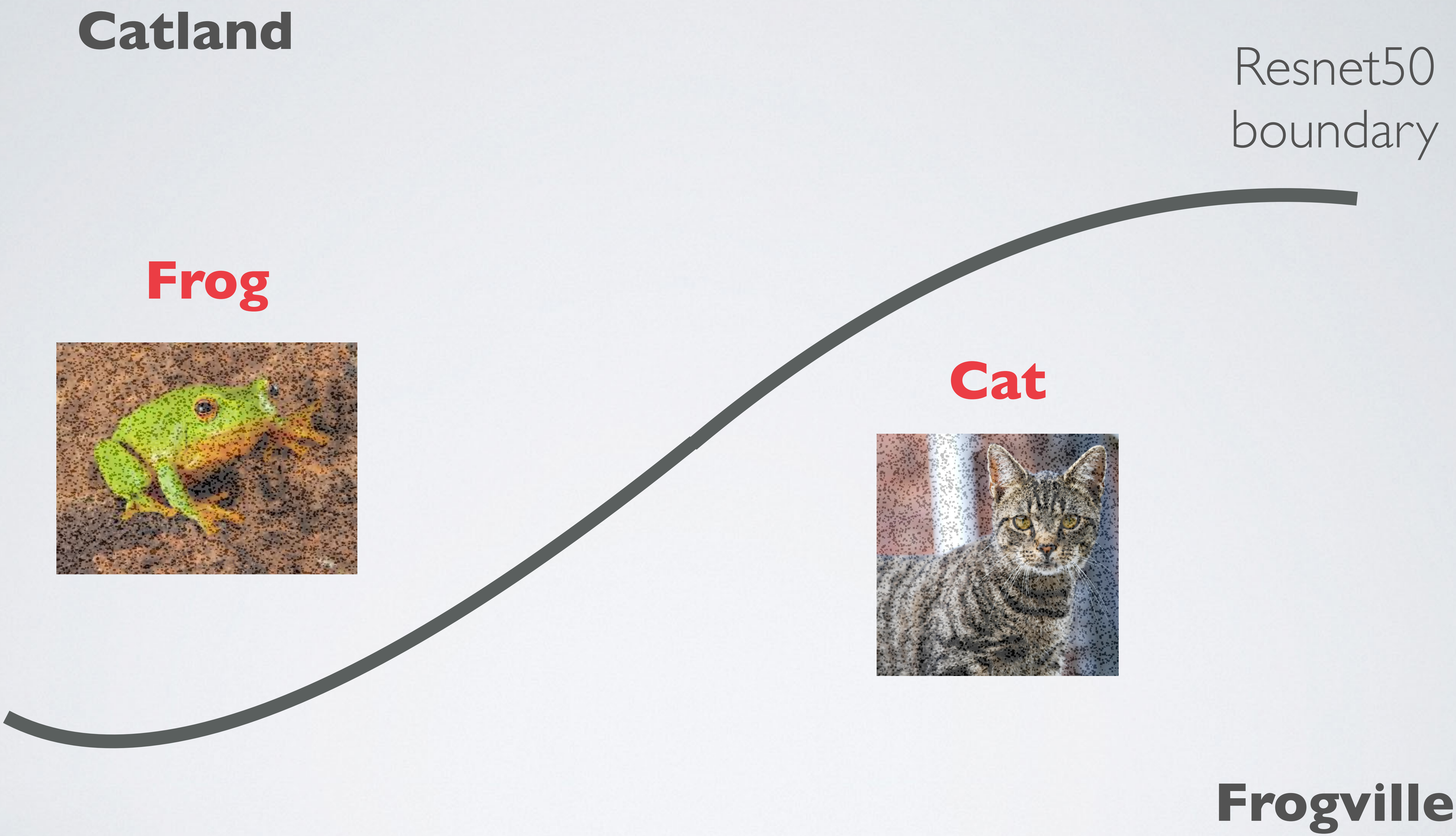
TRAIN ON ADVERSARIAL EXAMPLE TEST ON CLEAN DATA



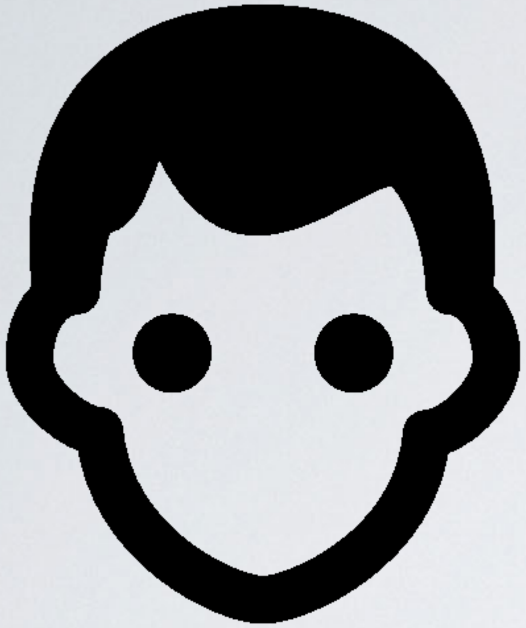
UNTRAINABLE IMAGENET?

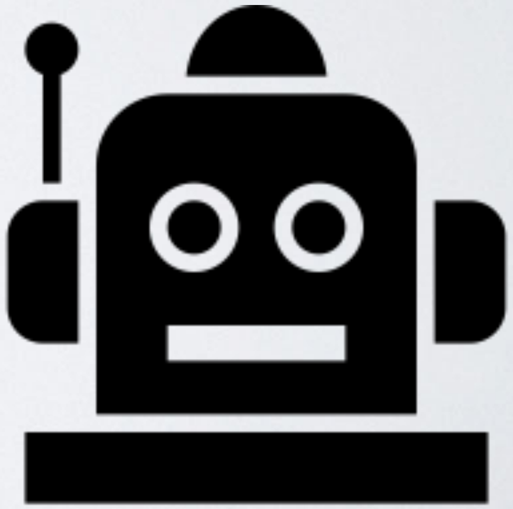
Images that are labeled "right" to a human but "wrong" to a computer.


Catland
Resnet50
boundary
Cat
Frog
Frogville


Catland
Resnet50
boundary
Frog
Cat
Frogville

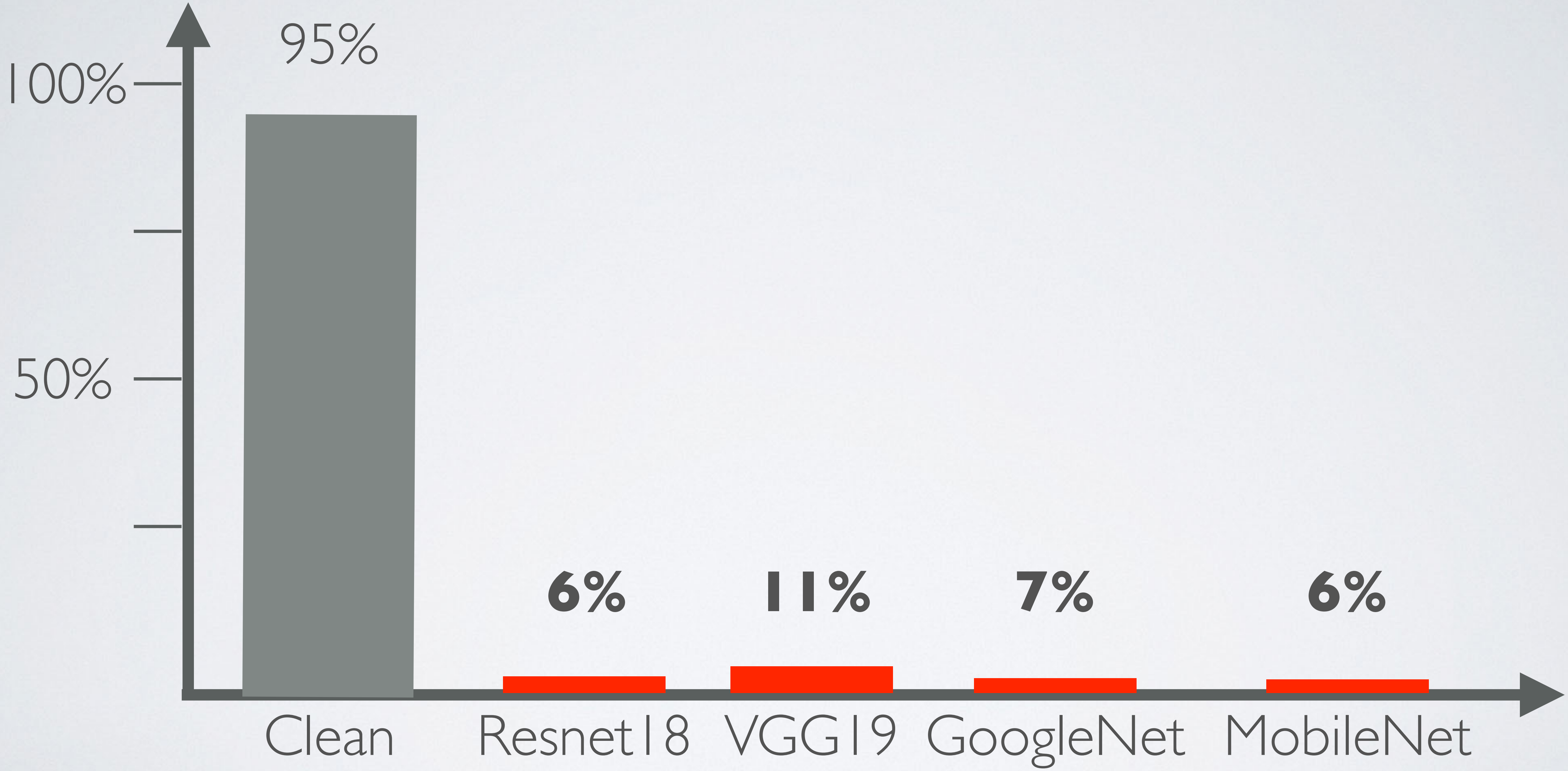
“Hen”





Ostrich

CIFAR-10 Accuracy
Poisoned with Resnet-18



"Adversarial examples make strong poisons"

Can you defeat poisoned data?

Adversarial Training

Table 4: **Adversarial Training.** CIFAR-10 test accuracy after adversarially training with different radii ρ_a. Top row shows performance of adversarial training on clean data. AR poisons remain effective for small ρ_a.

	ρ_a 0.125	0.25	0.50	0.75
Clean Data	87.07	84.75	81.19	77.01
• Error-Max [10]	$33.30_{\pm 0.14}$	$72.27_{\pm 2.18}$	81.15±[illegible].58	$78.73_{\pm 4.20}$
• Error-Min [18]	$70.66_{\pm 0.41}$	$84.80_{\pm 2.38}$	83.04±[illegible].24	$79.11_{\pm 3.46}$
∘ Regions-4	$75.05_{\pm 0.35}$	$81.23_{\pm 0.11}$	$79.71_{\pm 0.05}$	$76.47_{\pm 0.34}$
∘ Regions-16	$47.99_{\pm 0.25}$	$71.43_{\pm 0.17}$	[illegible]±0.10	$76.65_{\pm 0.07}$
∘ Random Noise	$86.31_{\pm 0.42}$	$84.17_{\pm 0.20}$	$\mathbf{80.11_{\pm 0.06}}$	$\mathbf{76.26_{\pm 0.07}}$
• Autoregressive (Ours)	$\mathbf{33.22_{\pm 0.77}}$	$\mathbf{57.08_{\pm 0.75}}$	$81.27_{\pm 2.61}$	$79.07_{\pm 3.47}$

Sandoval-Segura & G, *Autoregressive Perturbations*, 2022

Can you defeat poisoned data?

Mix with clean data

Table 5: **Mixing Poisons with Clean Data.** CIFAR-10 test accuracy when a proportion of clean data is used in addition to a poison. Top row shows test accuracy when training on only the clean proportion of the data; *i.e.* no poisoned data is used.

	Clean Proportion				
	40%	30%	20%	10%	5%
Clean Only	90.84	89.92	87.90	81.01	74.97
• Error-Max [18]	$87.83_{\pm 0.74}$	$86.83_{\pm 0.48}$	$84.70_{\pm 0.61}$	$81.63_{\pm 0.63}$	$76.48_{\pm 1.72}$
• Error-Min [10]	$88.32_{\pm 1.57}$	$87.23_{\pm 0.84}$	$84.56_{\pm 0.88}$	$78.76_{\pm 1.83}$	$67.82_{\pm 1.92}$
○ Regions-4	$88.94_{\pm 0.85}$	$86.75_{\pm 0.86}$	$83.52_{\pm 0.20}$	$78.23_{\pm 0.97}$	$70.19_{\pm 3.16}$
○ Regions-16	$88.03_{\pm 0.57}$	$86.23_{\pm 0.68}$	$\mathbf{83.01_{\pm 0.48}}$	$76.52_{\pm 0.91}$	$67.24_{\pm 1.72}$
○ Random Noise	$\mathbf{86.40_{\pm 1.24}}$	$86.99_{\pm 0.19}$	$84.98_{\pm 1.85}$	$78.08_{\pm 0.94}$	$70.69_{\pm 0.87}$
• AR (Ours)	$87.63_{\pm 0.68}$	$\mathbf{85.62_{\pm 0.62}}$	$83.28_{\pm 0.90}$	$\mathbf{76.13_{\pm 2.34}}$	$\mathbf{62.69_{\pm 5.58}}$

Sandoval-Segura & G, *Autoregressive Perturbations*, 2022

Data security in federated learning

FEDERATED LEARNING

GBoard Predictive text

ML Kit

Image recognition API

App monitoring
& marketing data

GOING BEYOND PATTERN MATCHING

WHAT'S FEDERATED LEARNING?

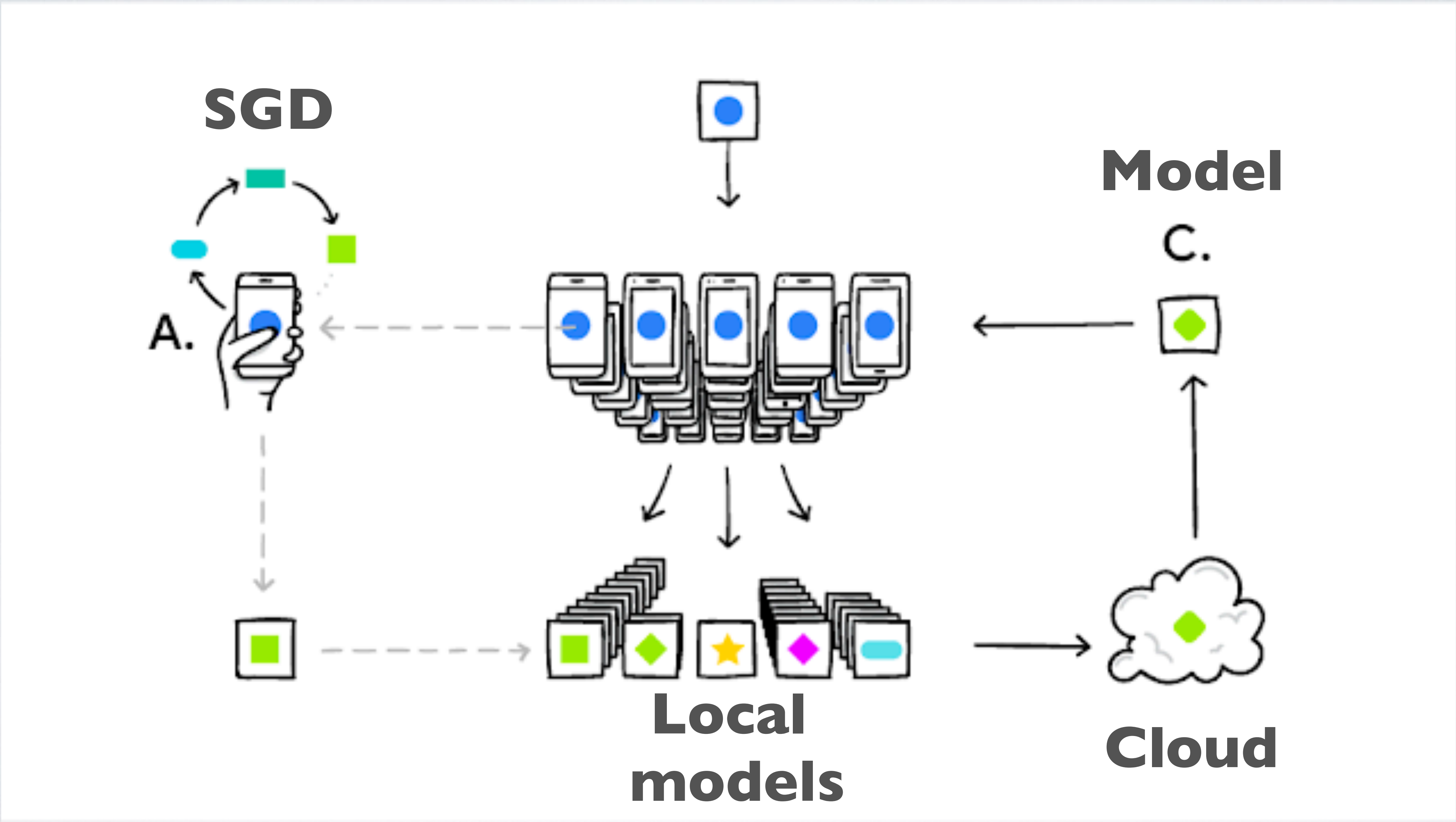


Figure stolen from ai.googleblog.com/2017/04/federated-learning-collaborative.html

IS IT PRIVATE?

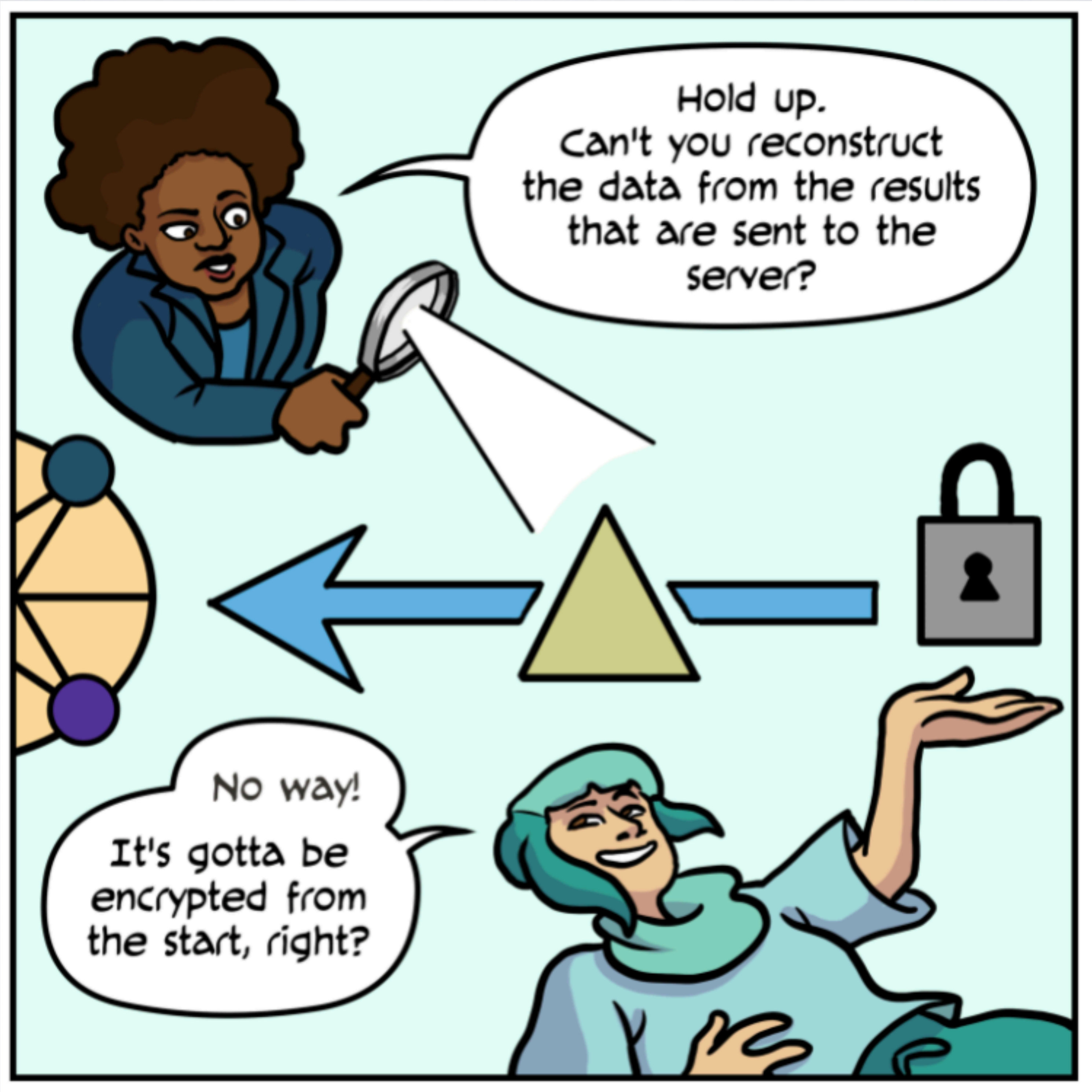


Figure stolen from https://federated.withgoogle.com/

A BIG LEAK

BIG SECURITY LEAK: LINEAR LAYERS

Linear layers

$$z = Wx + b$$

Downstream loss

$$\mathcal{L}(z)$$

Parameter gradients

$$\nabla_W \mathcal{L} = \nabla_z \mathcal{L}(z) x$$

$$\nabla_b \mathcal{L} = \nabla_z \mathcal{L}(z)$$

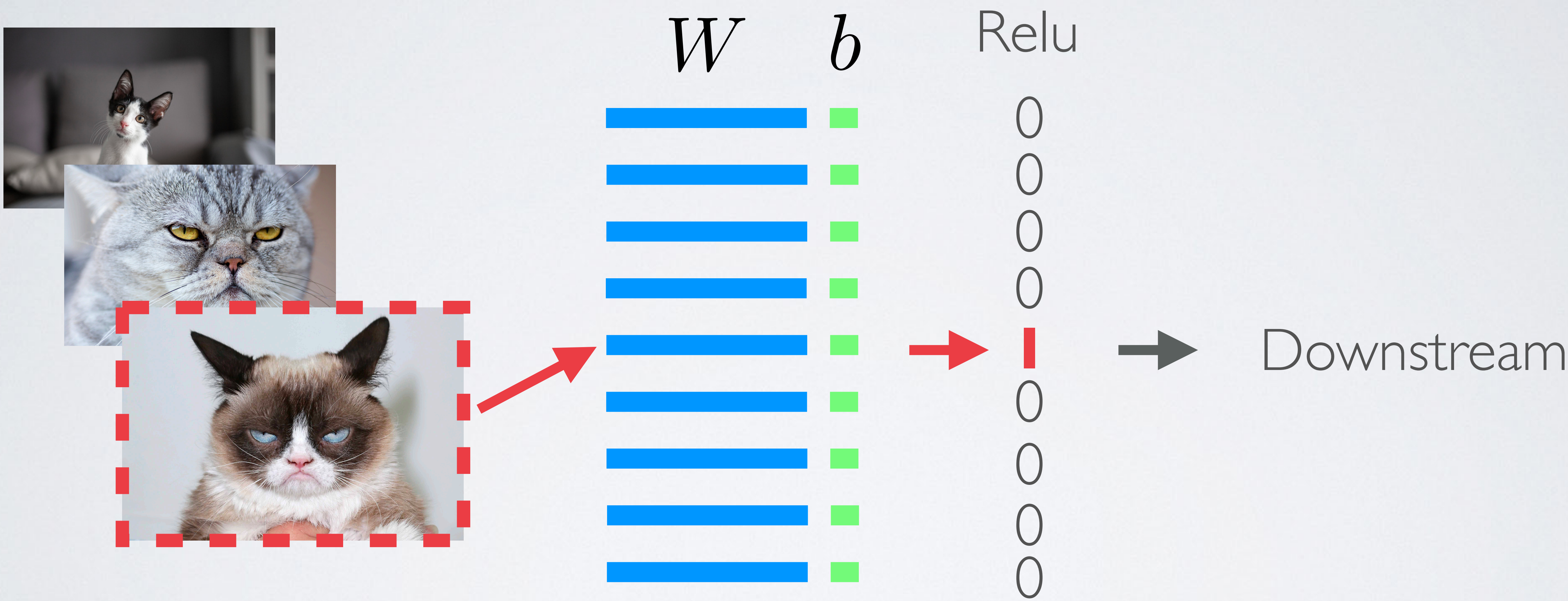
Uh oh.

$$x = \nabla_W \mathcal{L} / \nabla_b \mathcal{L}$$

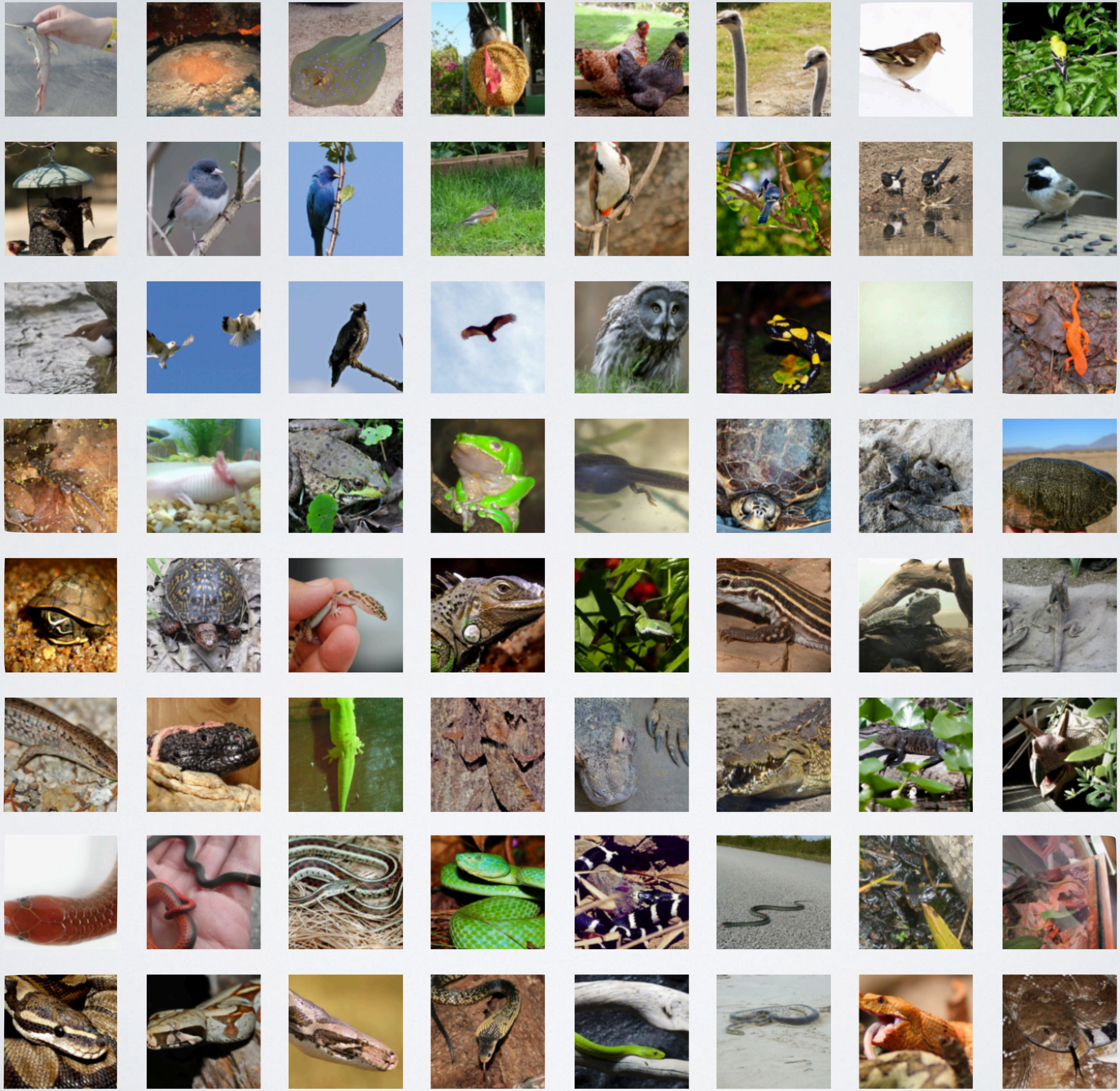
Fowl et al. "Robbing the Fed." 2021

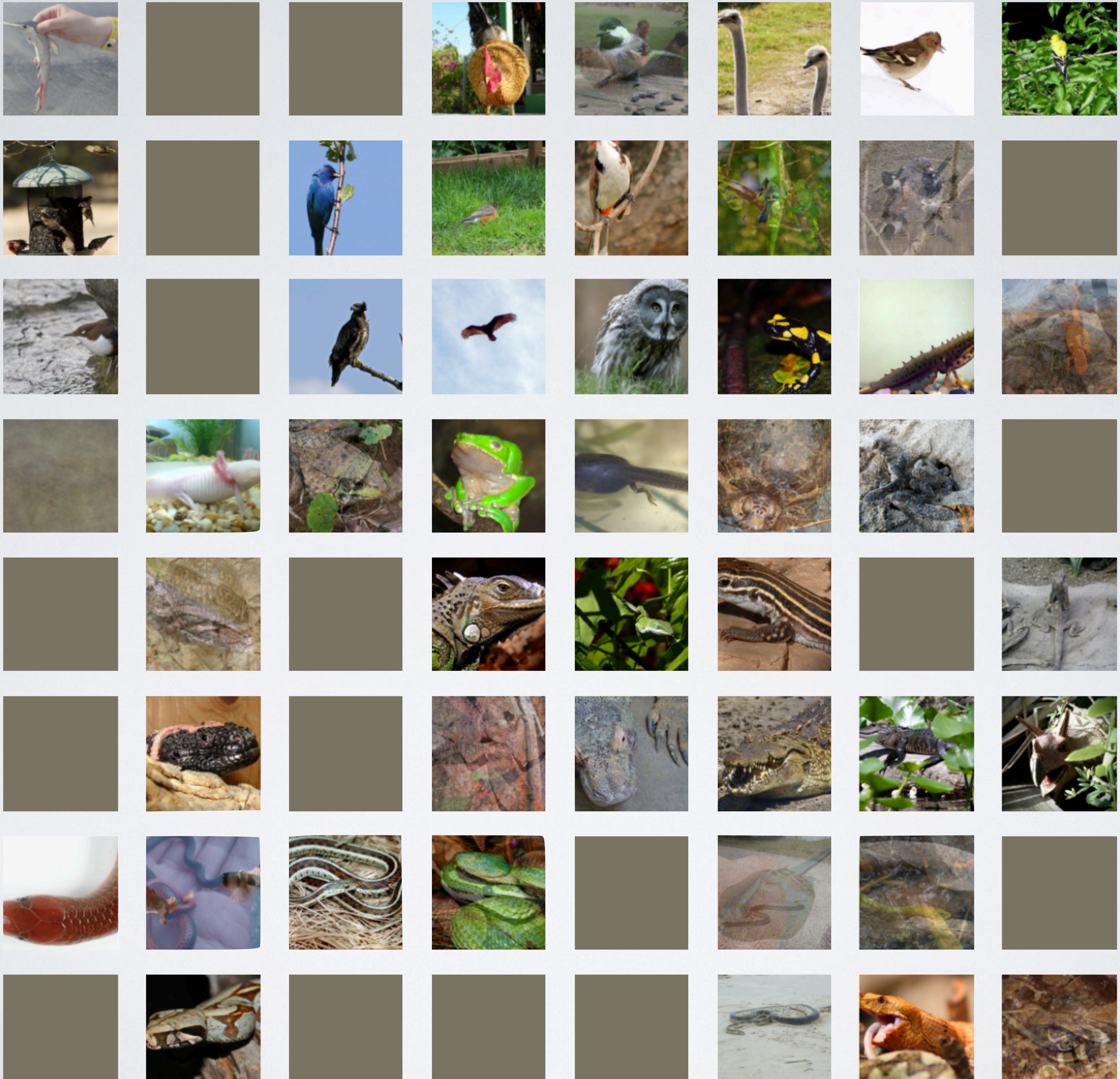
BUT WE'RE PROTECTED BY BATCHING! …RIGHT?

Linear layer filters



Fowl et al. "Robbing the Fed." 2021





EXAMPLE

batch size 16K

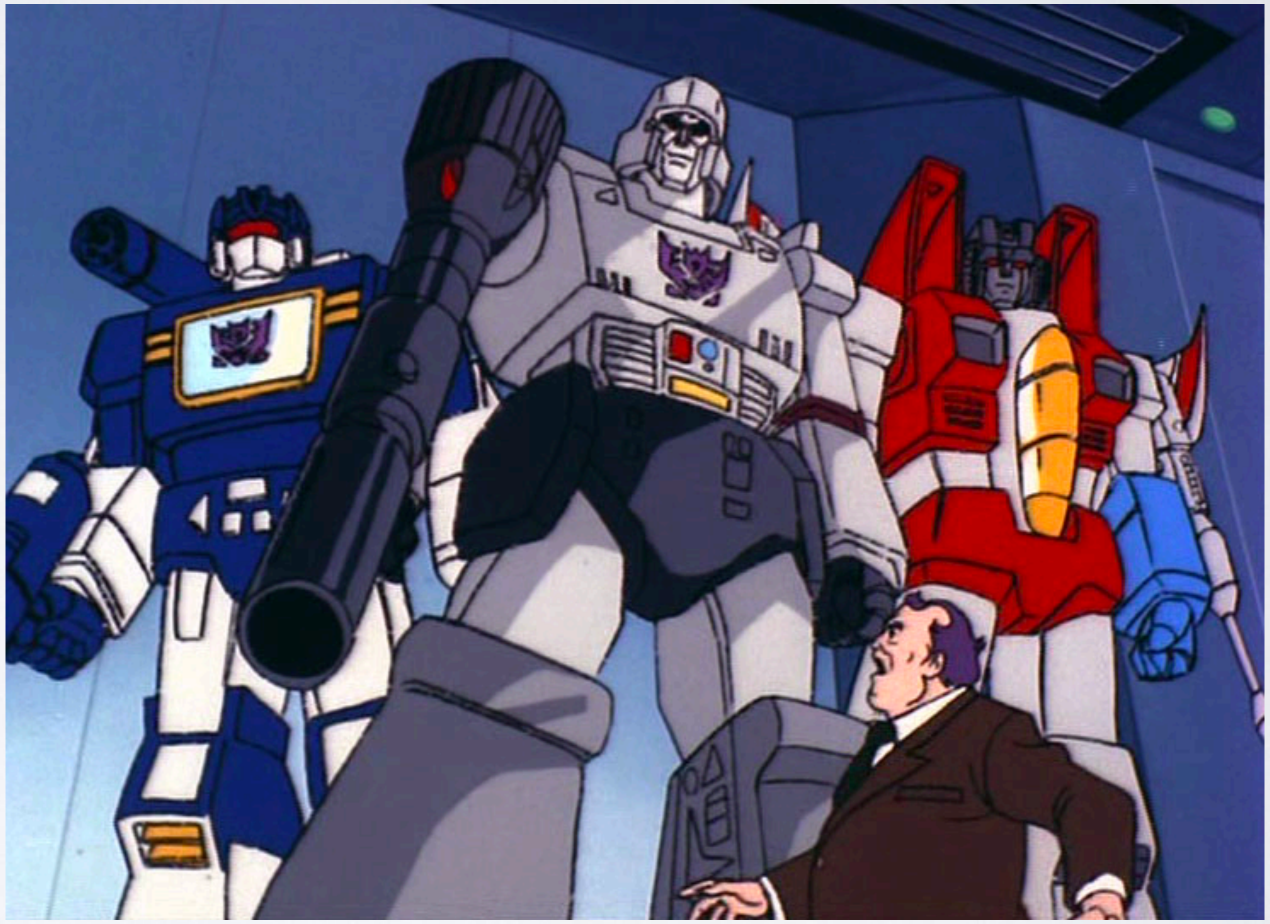
Original

Imprinted

Fowl et al. “Robbing the Fed.” 2021

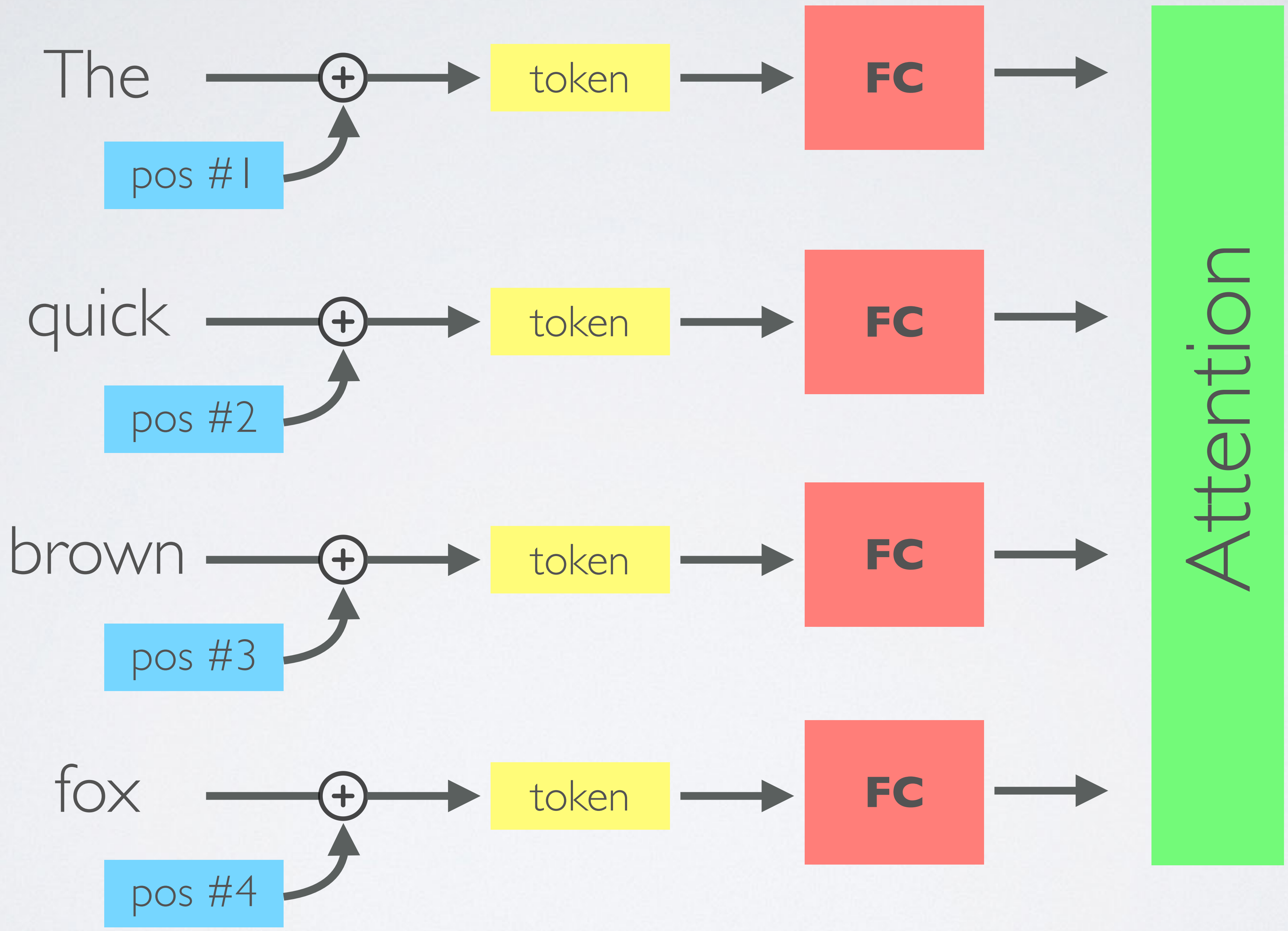
But what about text?

Decepticons

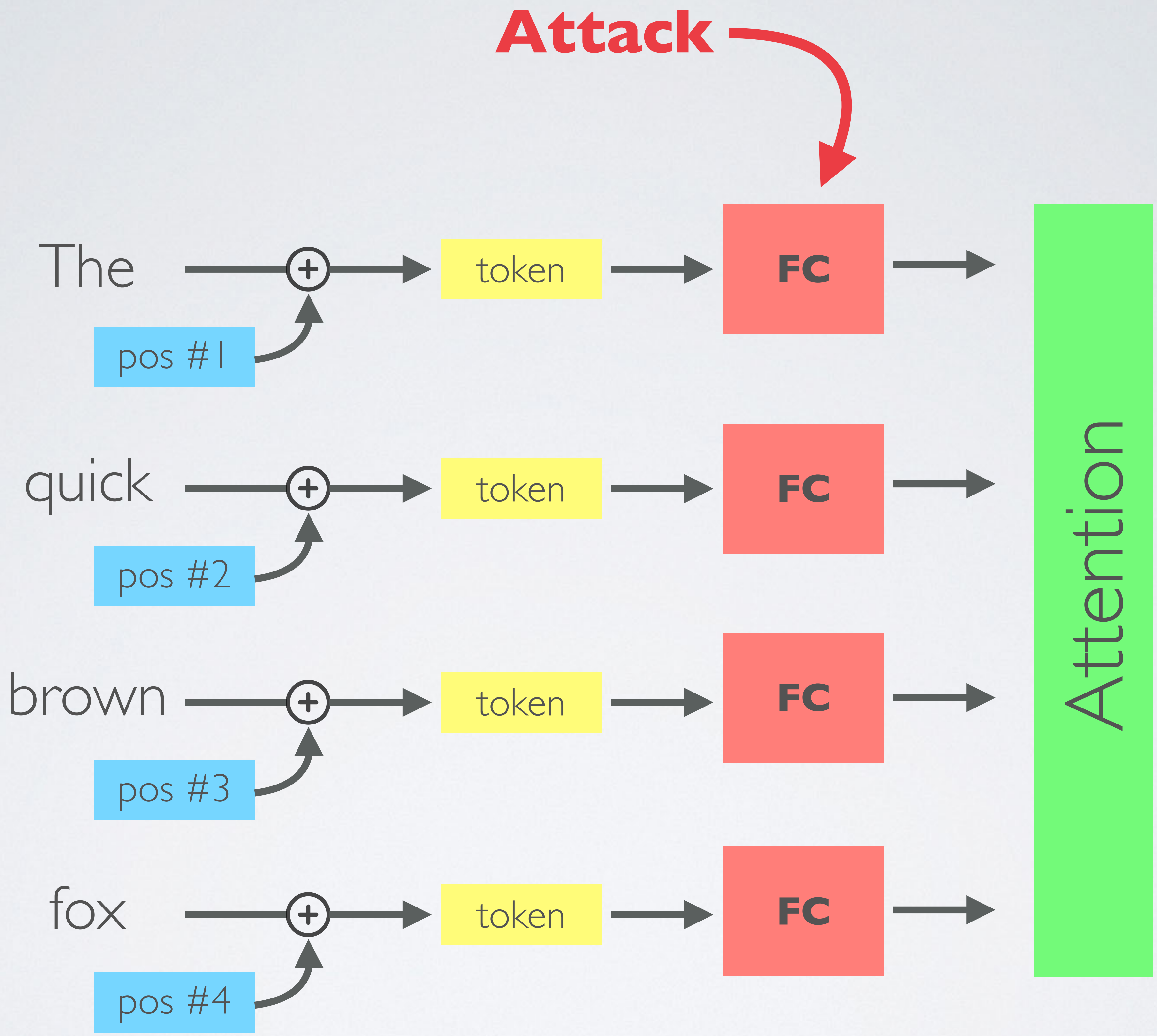


“Decepticons: Corrupted Transformers Breach Privacy in Federated Learning for Language Models”

Text transformers



“Decepticons: Corrupted Transformers Breach Privacy in Federated Learning for Language Models”

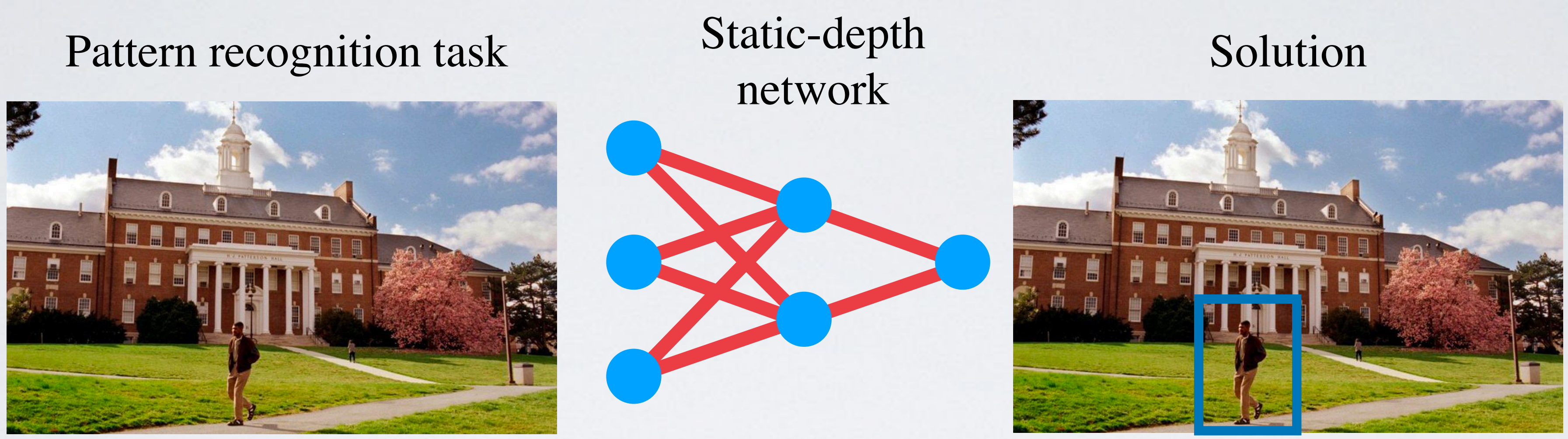


“Decepticons: Corrupted Transformers Breach Privacy in Federated Learning for Language Models”

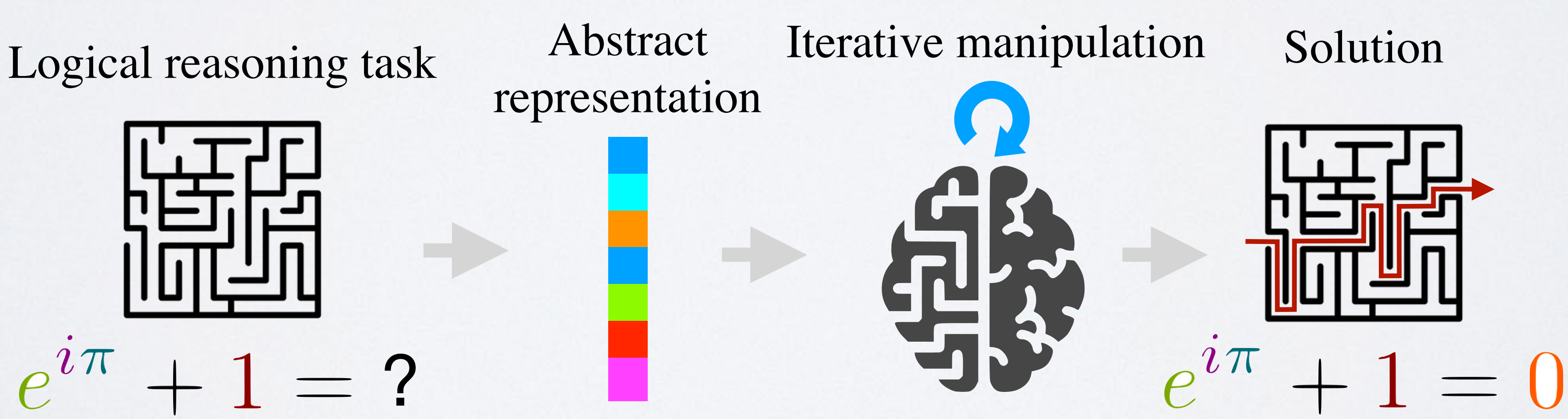
	Batch Size = 1	Batch Size = 8	Batch Size = 16
Length 32	Ancient Egyptian deities Egypt the gods and goddesses worshipped. ancient gods are The beliefs of rituals surrounding these in	Ancient Egyptian deities are the gods and goddesses worshipped in ancient Egypt ph The beliefs and rituals surrounding these gods	Ancient for deities are the gods and goddesses worshipped in ancient Egypt. The beliefs and rituals surrounding these gods
Length 128	Ancient Egyptian deities are the gods and goddesses worshipped Egypt ancient constitu. The beliefs and rituals myths these gods	Ancient Egyptian deities are the gods and goddesses worshipped in ancient Egypt. The beliefs view rituals surrounding these gods	Ancient Egyptian deities are the gods and goddesses worshipped in ancient Egypt. The beliefs view rituals surrounding these continue
Length 512	Ancient Egyptian well are the gods and goddesses worshipped in ancient Egypt � The beliefs whereas ritualsies these gods formed	Ancient Egyptian deities are the gods and goddesses worshipped in ancient vague. " beliefs and. tried these gods	Ancient Egyptian deities are the gods and goddess hours thoughts in ancient final conception divine beliefs and rituals and these

Building and breaking thinking systems

“Fast”/Type-I thinking

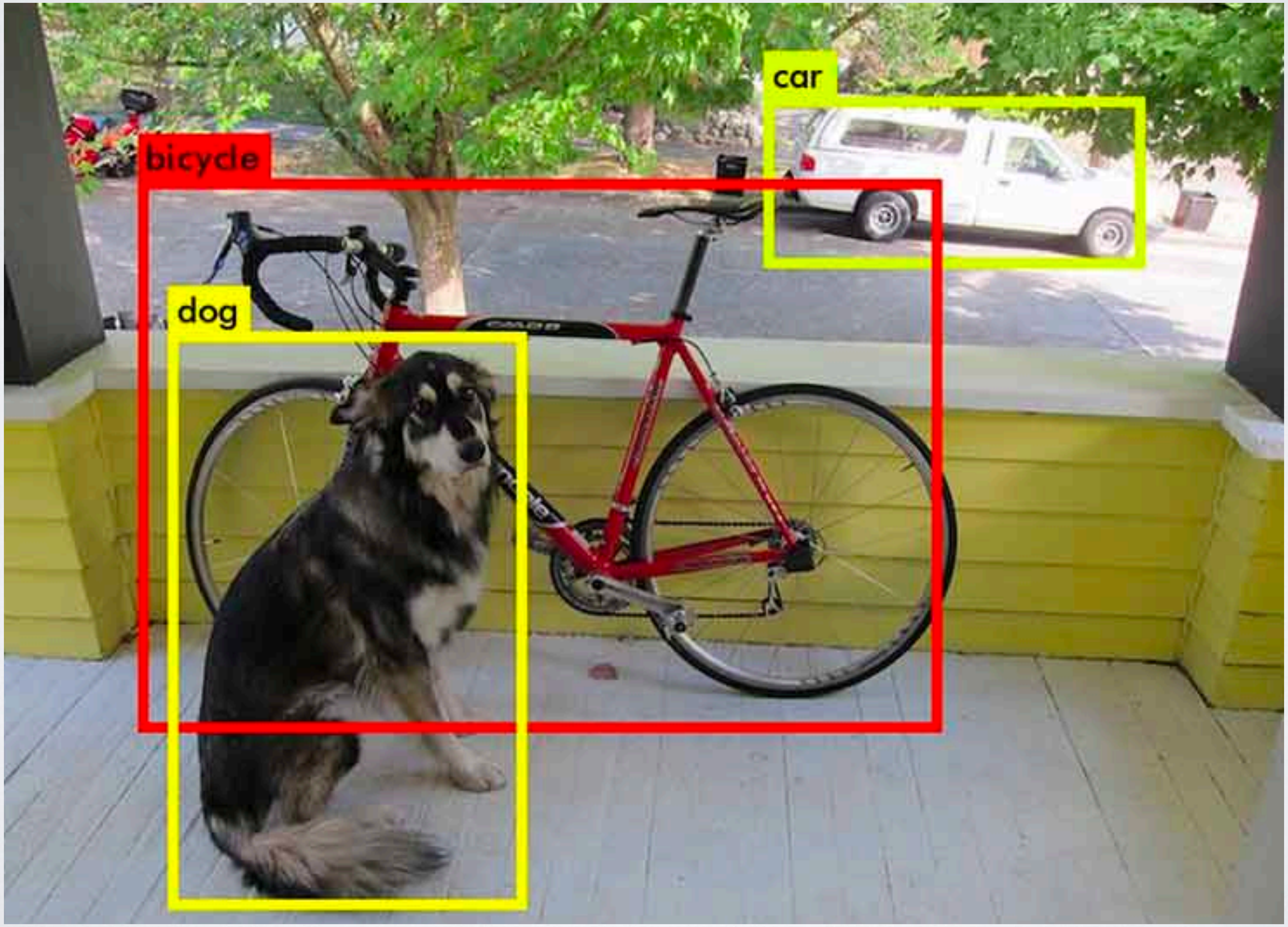


“Slow”/Type-II thinking



Machines are better than humans at...

Pattern matching
"Type 1 thinking"



Type II thinking = logical reasoning

Human reasoning scales to problems of (potentially) unbounded difficulty

$$e^{i\pi}+1=0$$

Humans handle domain shift well

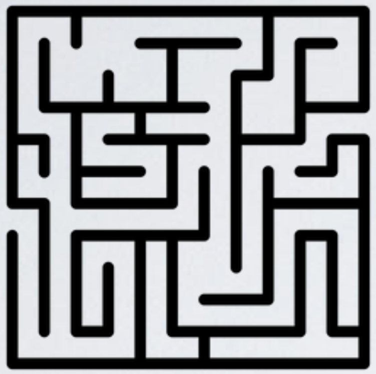
Humans can synthesize complex strategies from simple rules

Can neural networks exhibit logical extrapolation?

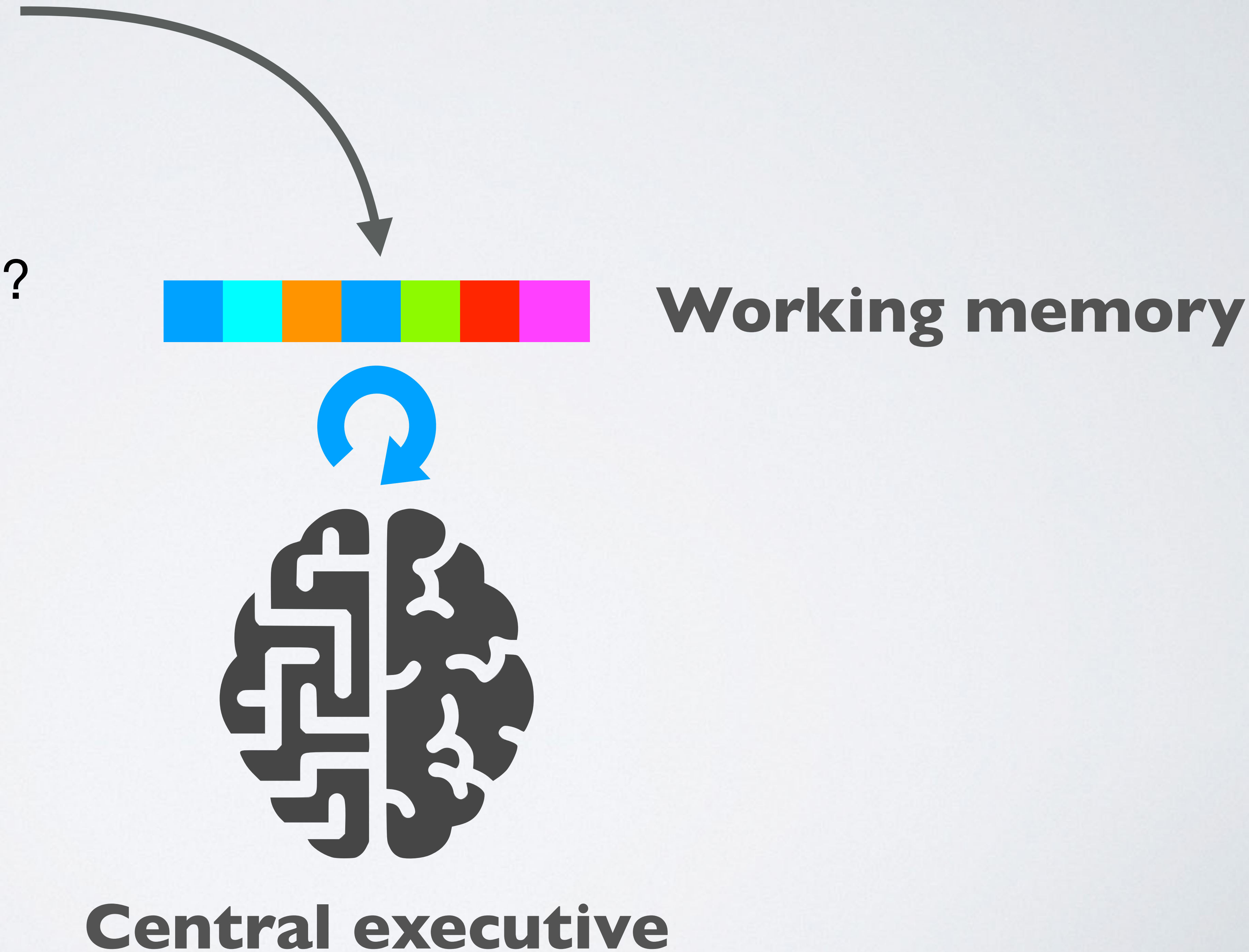
I.e., a system that solves problems of unlimited complexity just by "thinking for longer?"

Why can humans perform logical extrapolation?

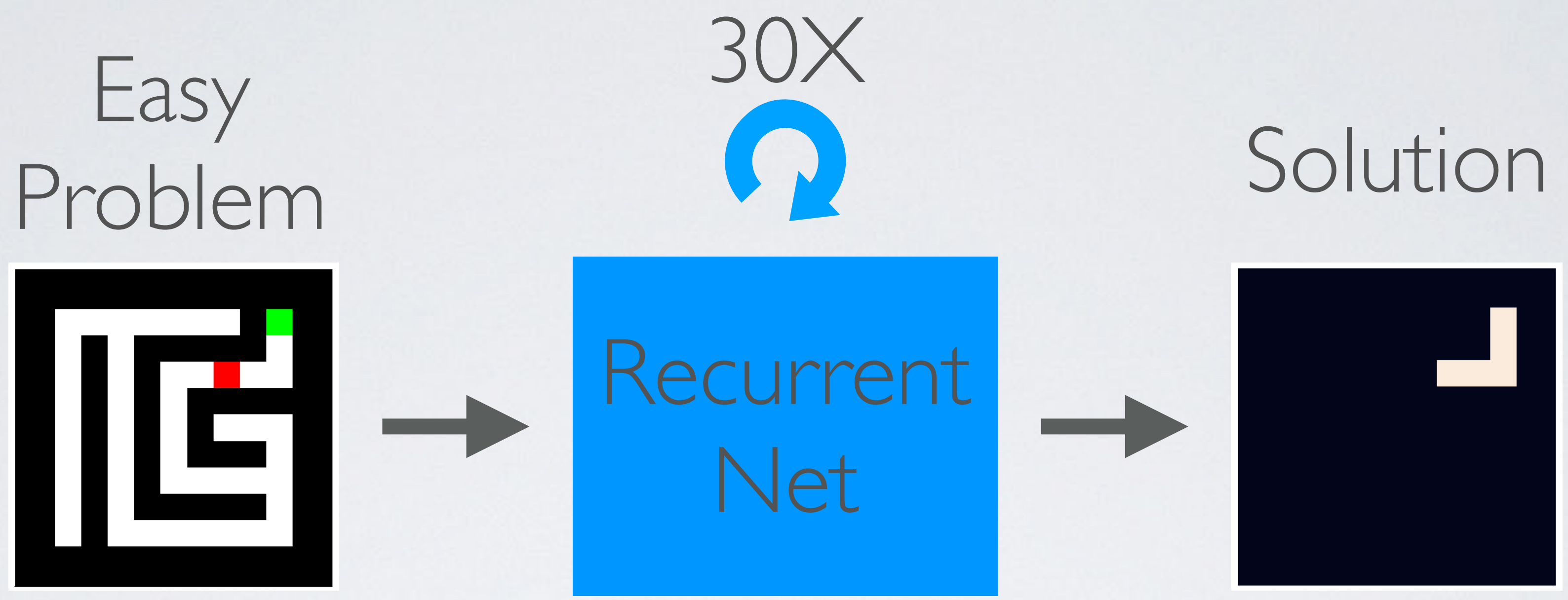
Logical reasoning task



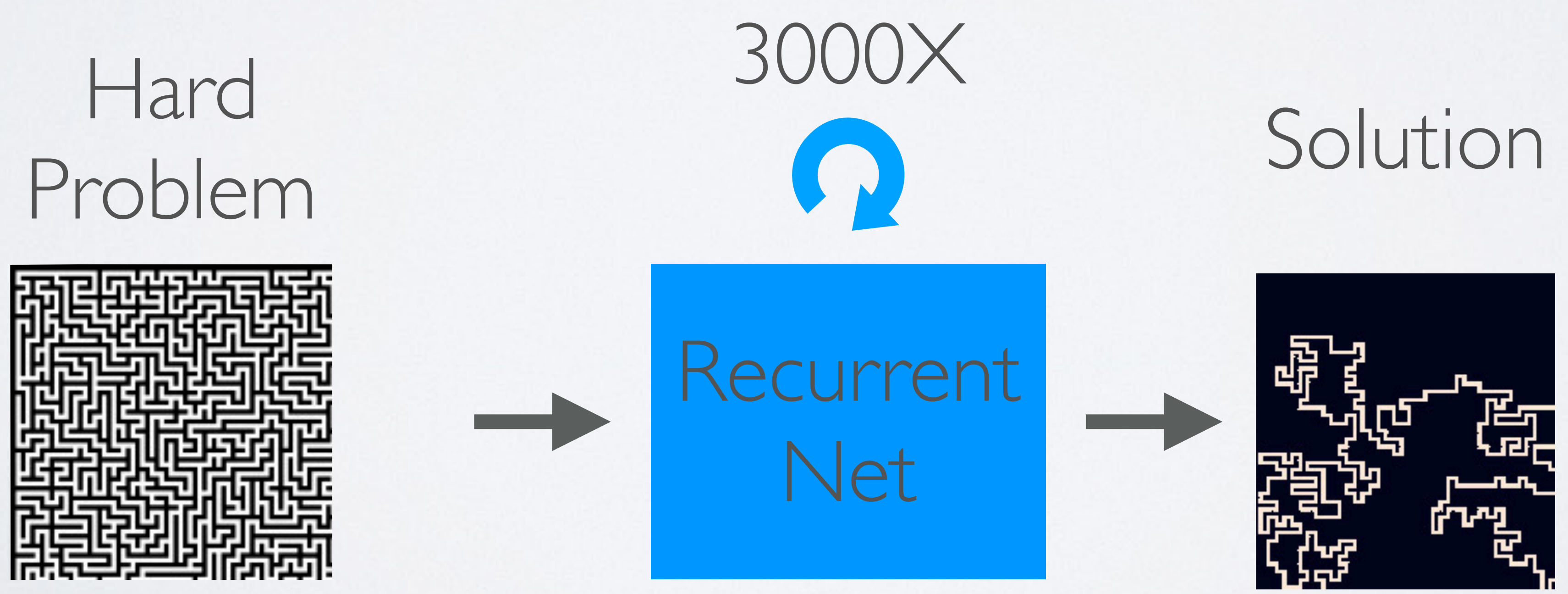
$$e^{i\pi} + 1 = \ ?$$



Train

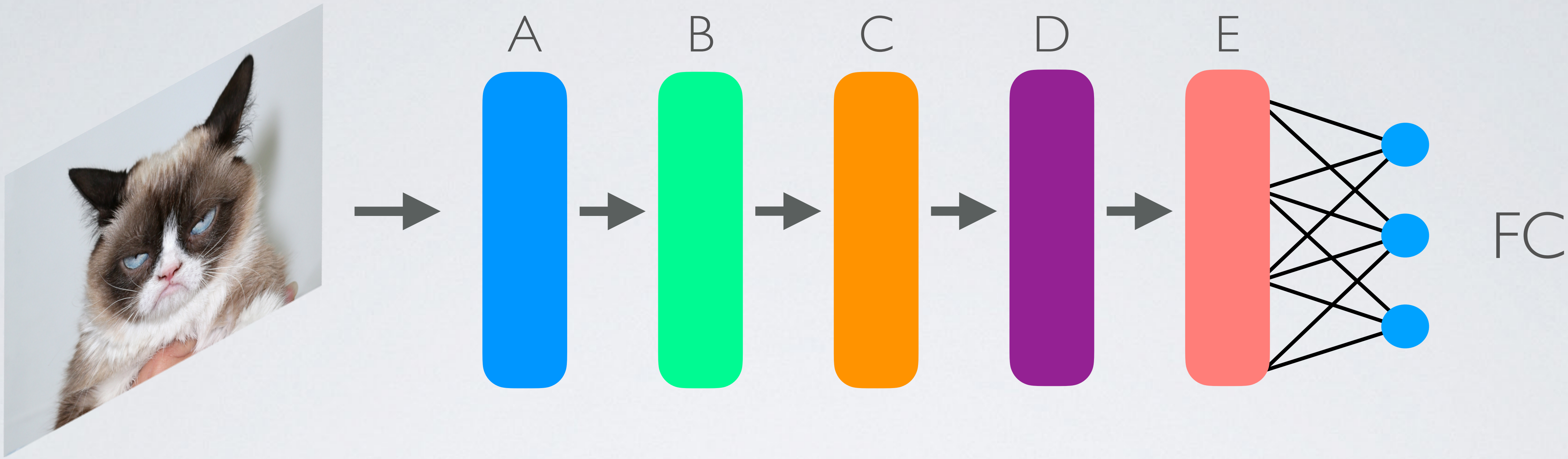

Easy
Problem
30X
Recurrent
Net
Solution

Test

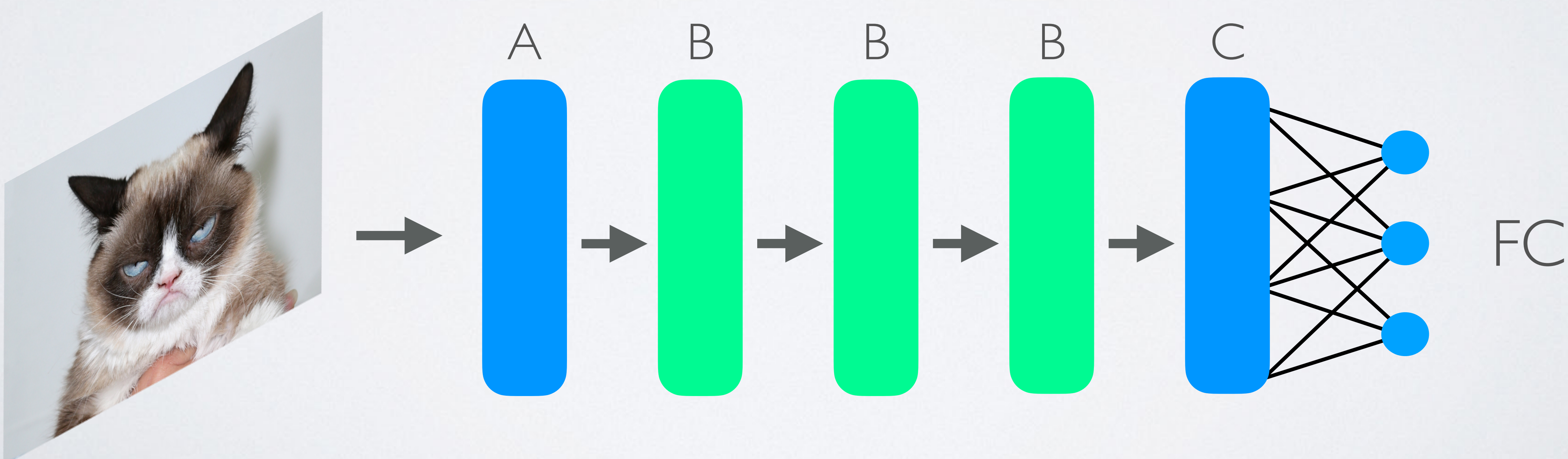

Hard
Problem
3000X
Recurrent
Net
Solution

Getting started: Replace feed-forward computation with recurrence

Feedforward model

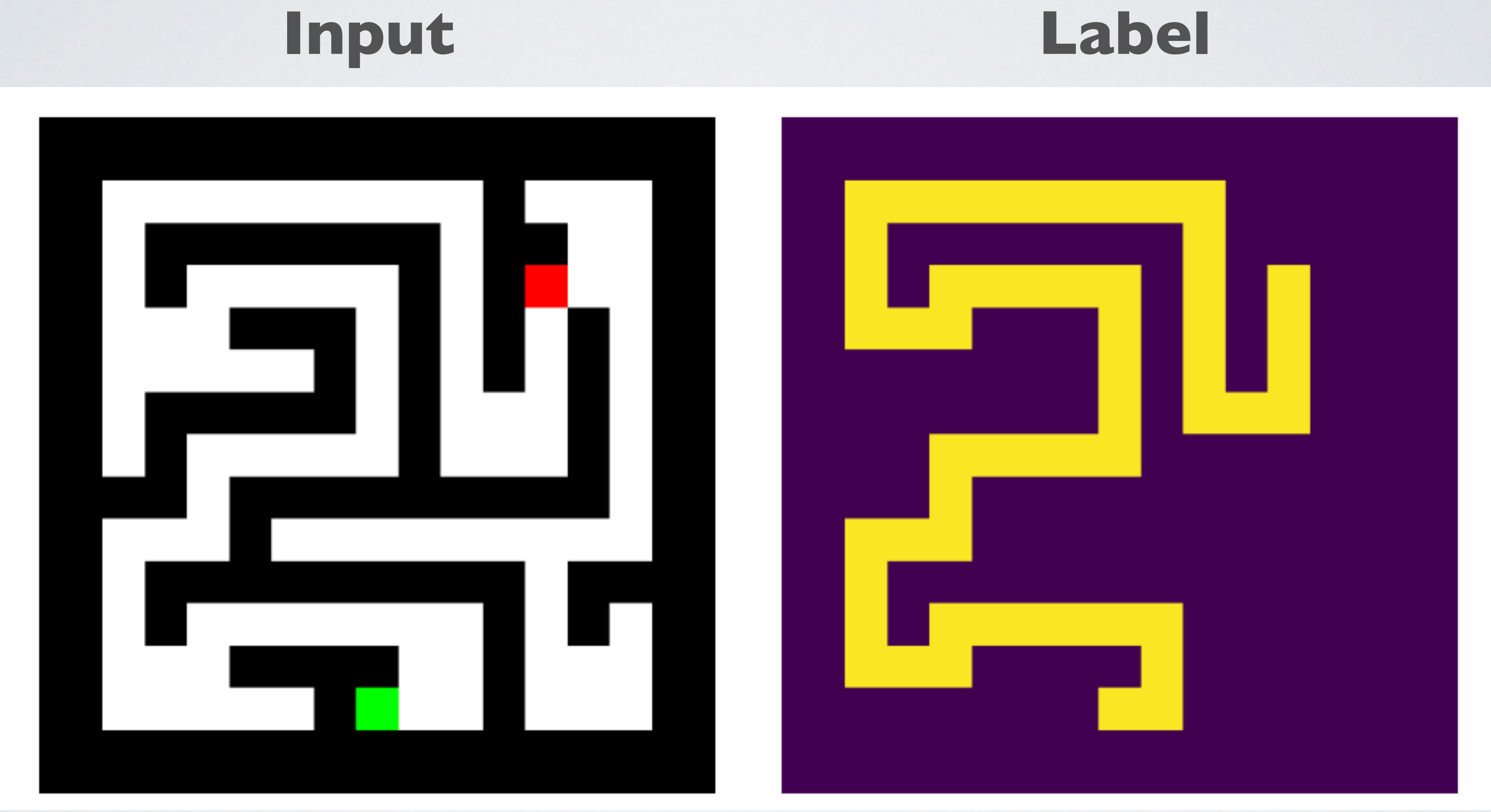


Recurrent model



Controlling the hardness of a problem

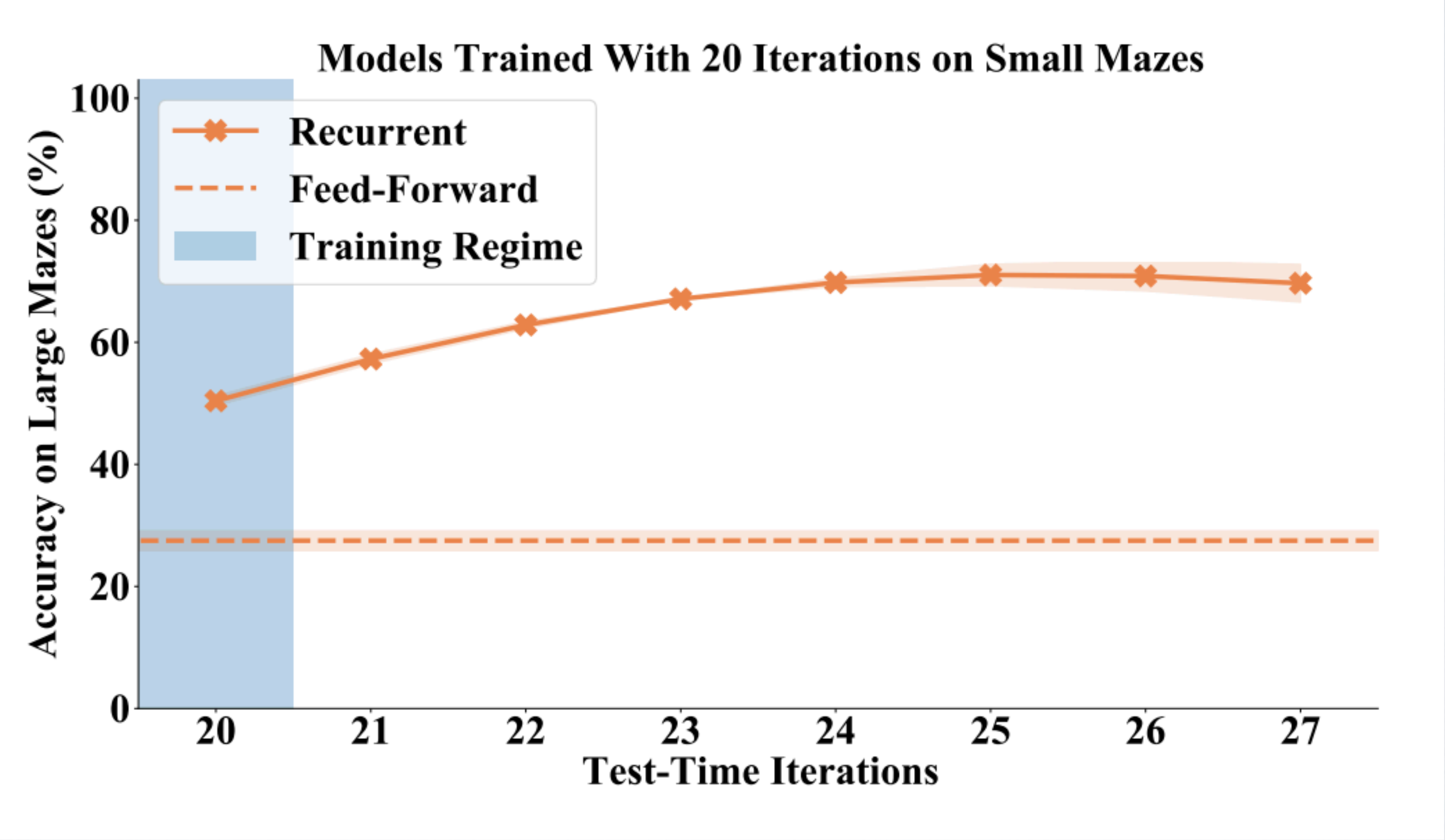
Procedurally generated mazes



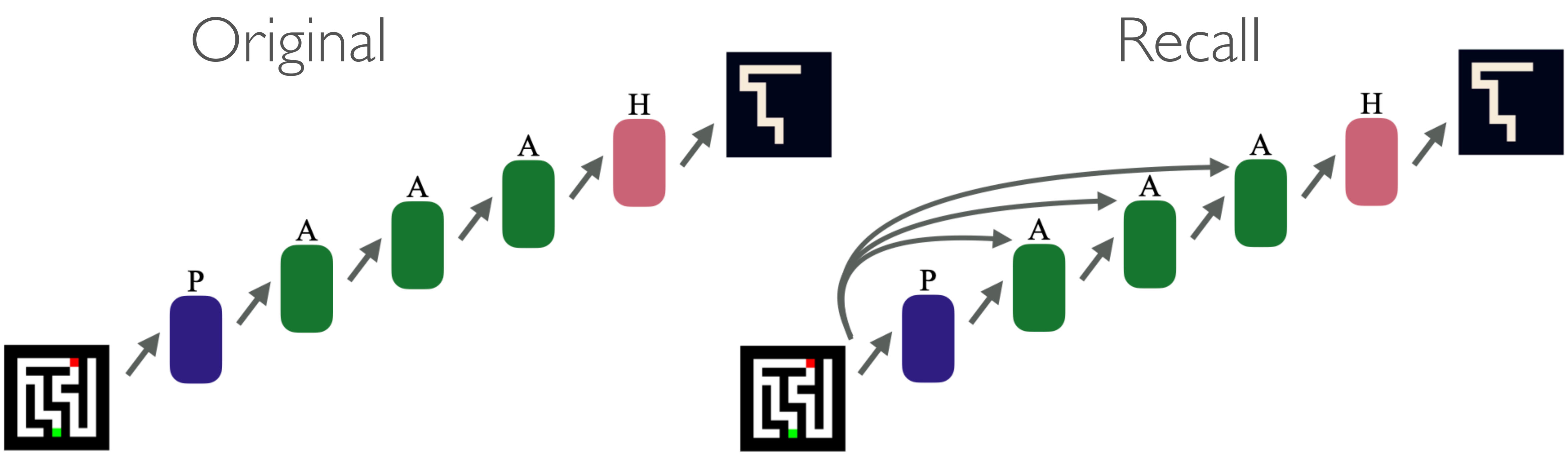
Schwarzschild et al. "Datasets for Studying Generalization from Easy to Hard Examples"

MAZES

Train on 9x9, test on 13x13

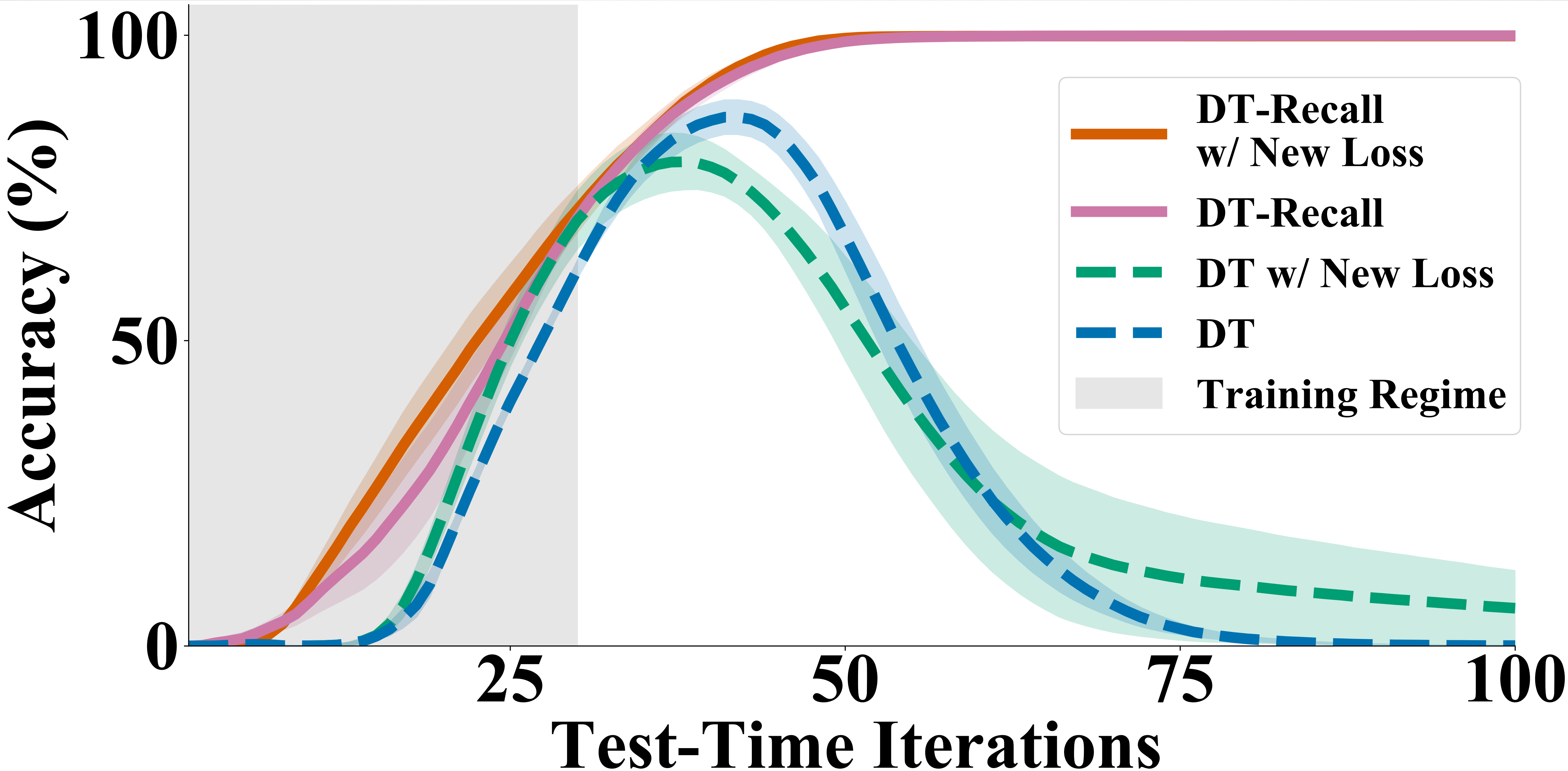


ARCHITECTURE IMPROVEMENT



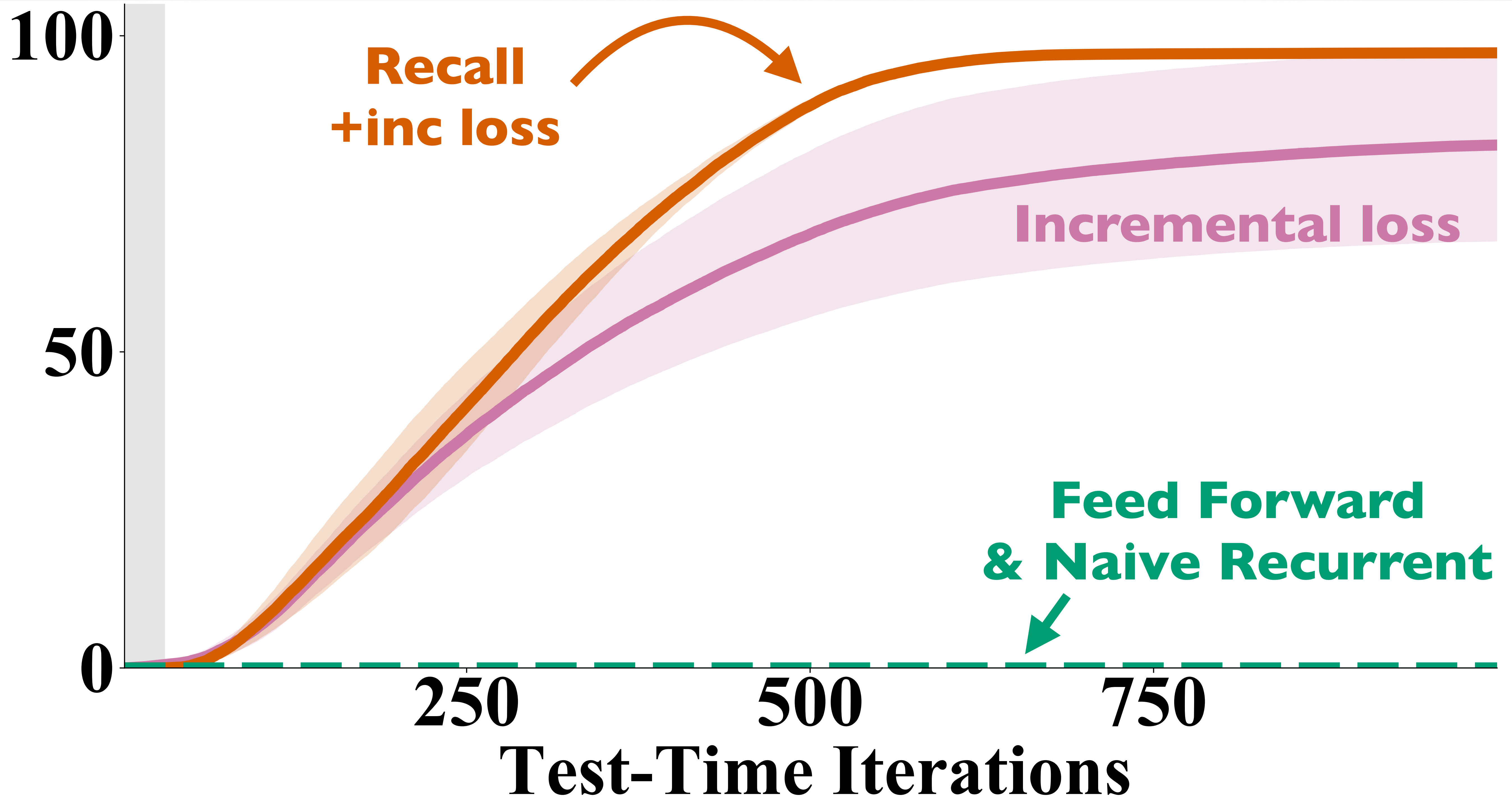
“Thinking Deeper with Recurrence,” 21

Train on 9x9 → Test on 13x13

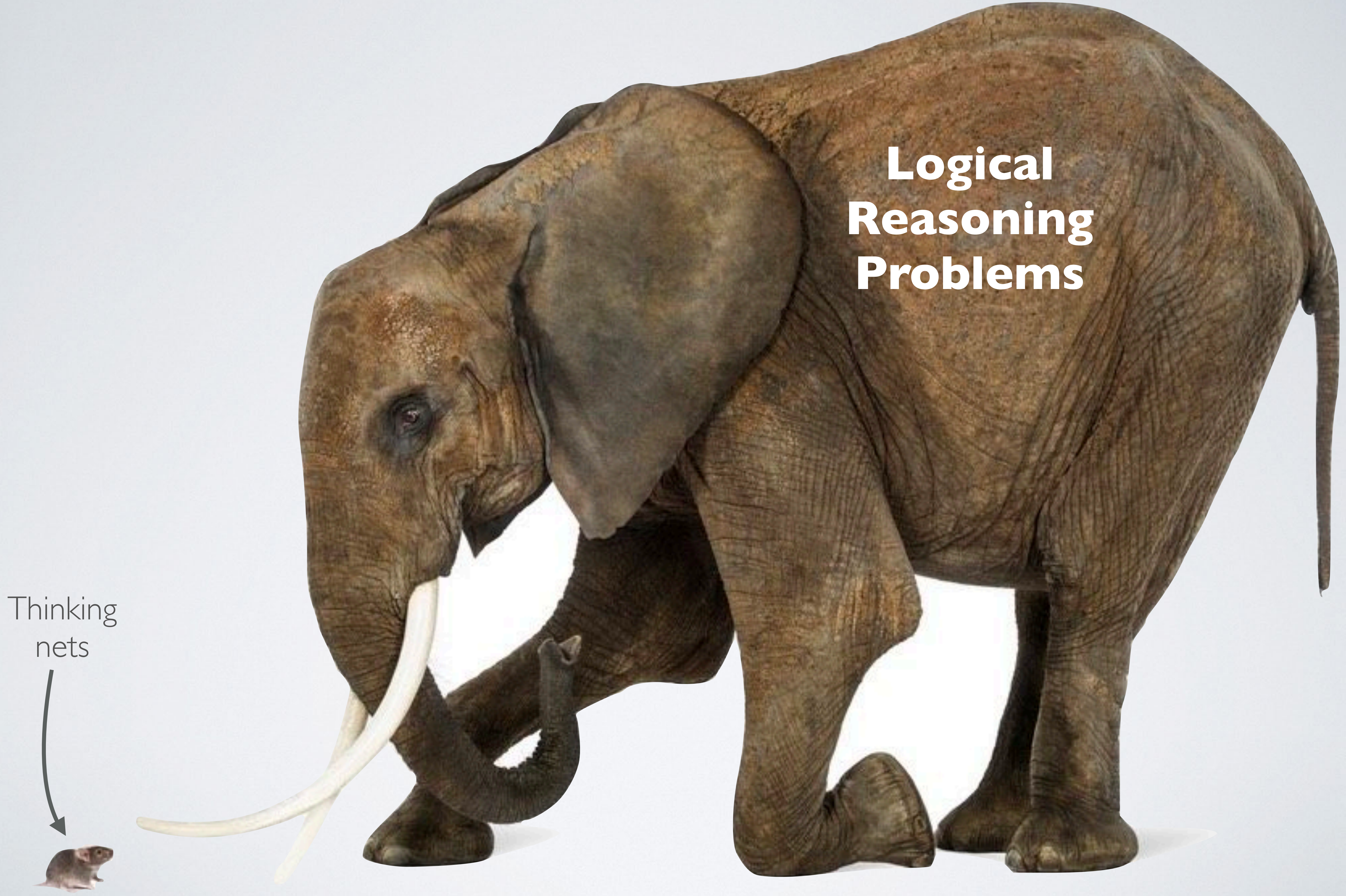


"Thinking Deeper with Recurrence," 21

Train on 9x9 → Test on 59x59



SCALING UP



A problem that can be solved by a simple “for” loop

Test problem: Prefix sums

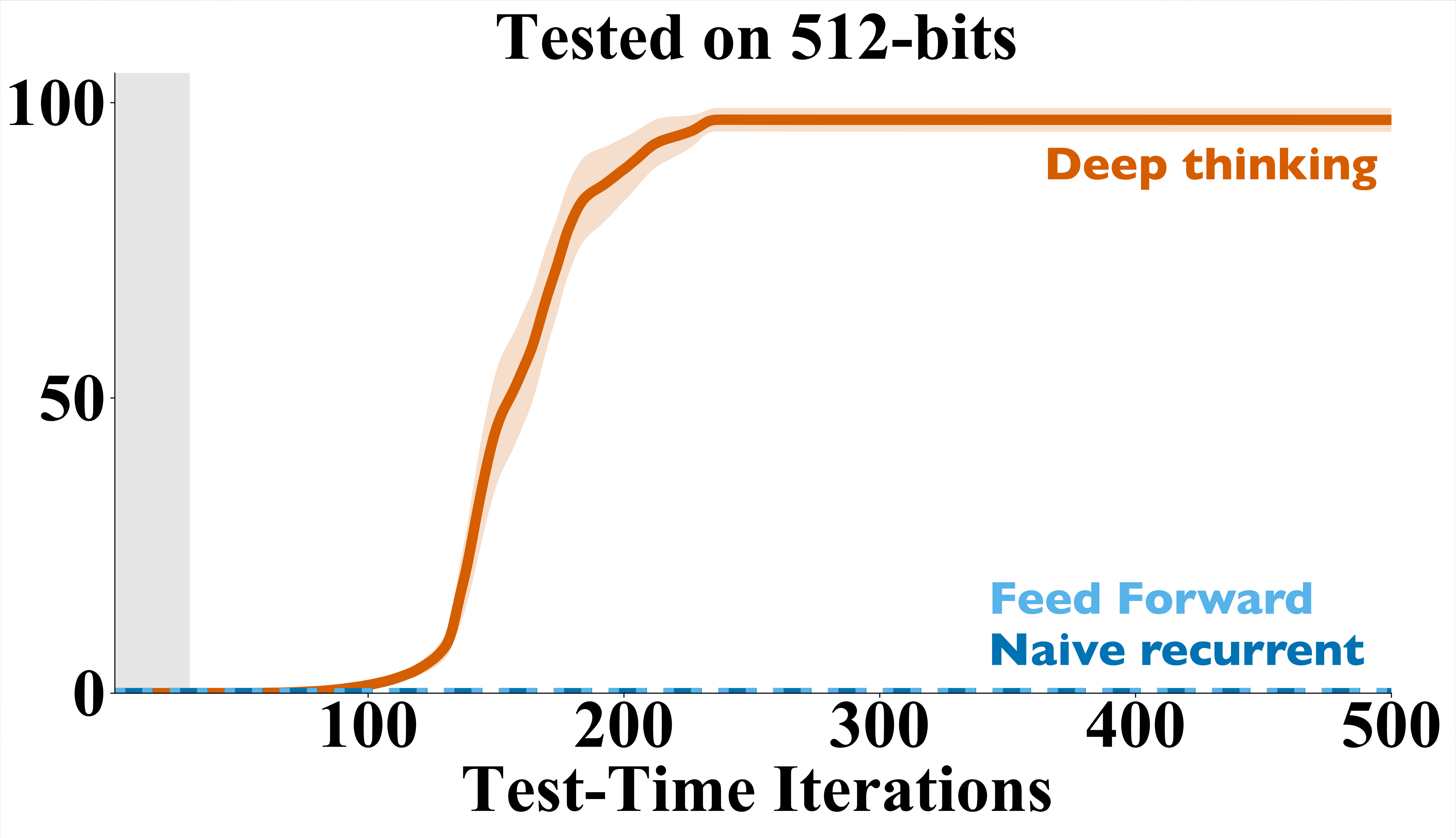
Goal: compute cumulative sum mod 2

Input:	$[1, 0, 1, 0, 1, 0, 1, 1, 0, 0, 1, 1, 1, 0, 1, 1]$
Target:	$[1, 1, 0, 0, 1, 1, 0, 1, 1, 1, 0, 1, 0, 0, 1, 0]$
Input:	$[1, 0, 0, 1, 1, 0, 1, 1, 0, 1, 1, 0, 1, 0, 1, 1, 0, 0, 0, 1, 1, 0, 1, 0, 1, 1, 0, 0]$
Target:	$[1, 1, 1, 0, 1, 1, 0, 1, 1, 0, 1, 1, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 1, 1, 0, 1, 1, 1]$

Schwarzschild et al. "Datasets for Studying Generalization from Easy to Hard Examples"

ARCHITECTURE IMPROVEMENT

Train on 32 bits

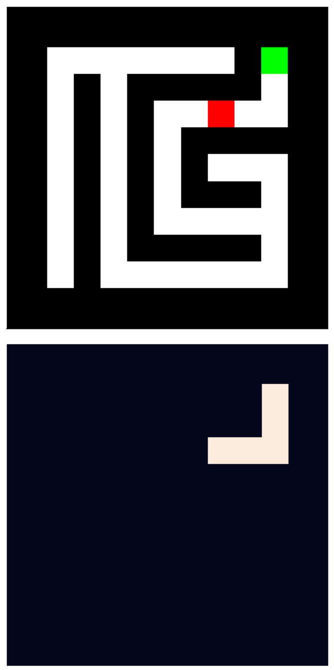


A problem that requires branching

Train on this.

30 iterations

9x9



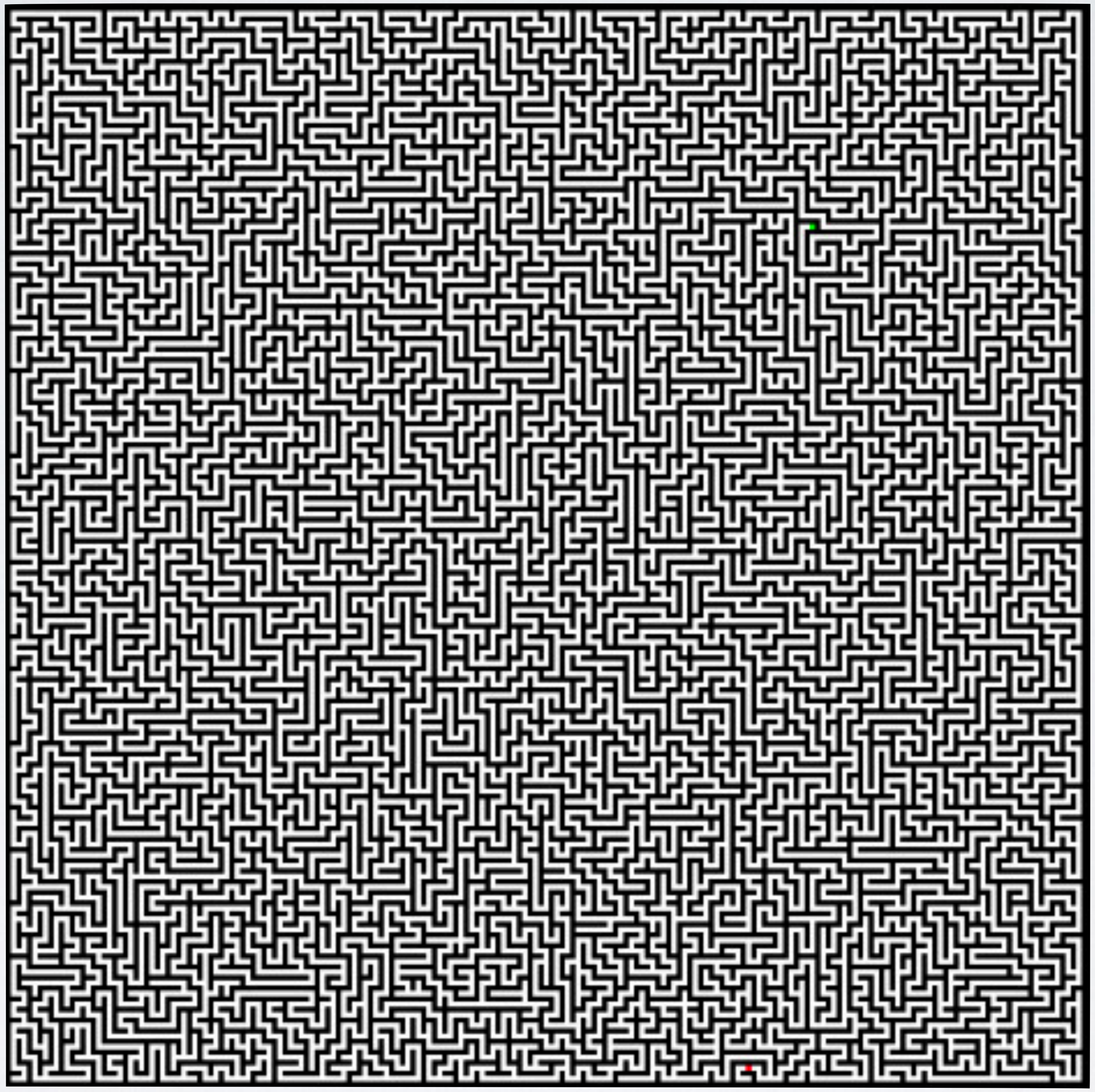
Schwarzschild et al. "Datasets for Studying Generalization from Easy to Hard Examples"

Train on 9x9 → Test on 201x201

30 iterations

2400 iterations

9x9

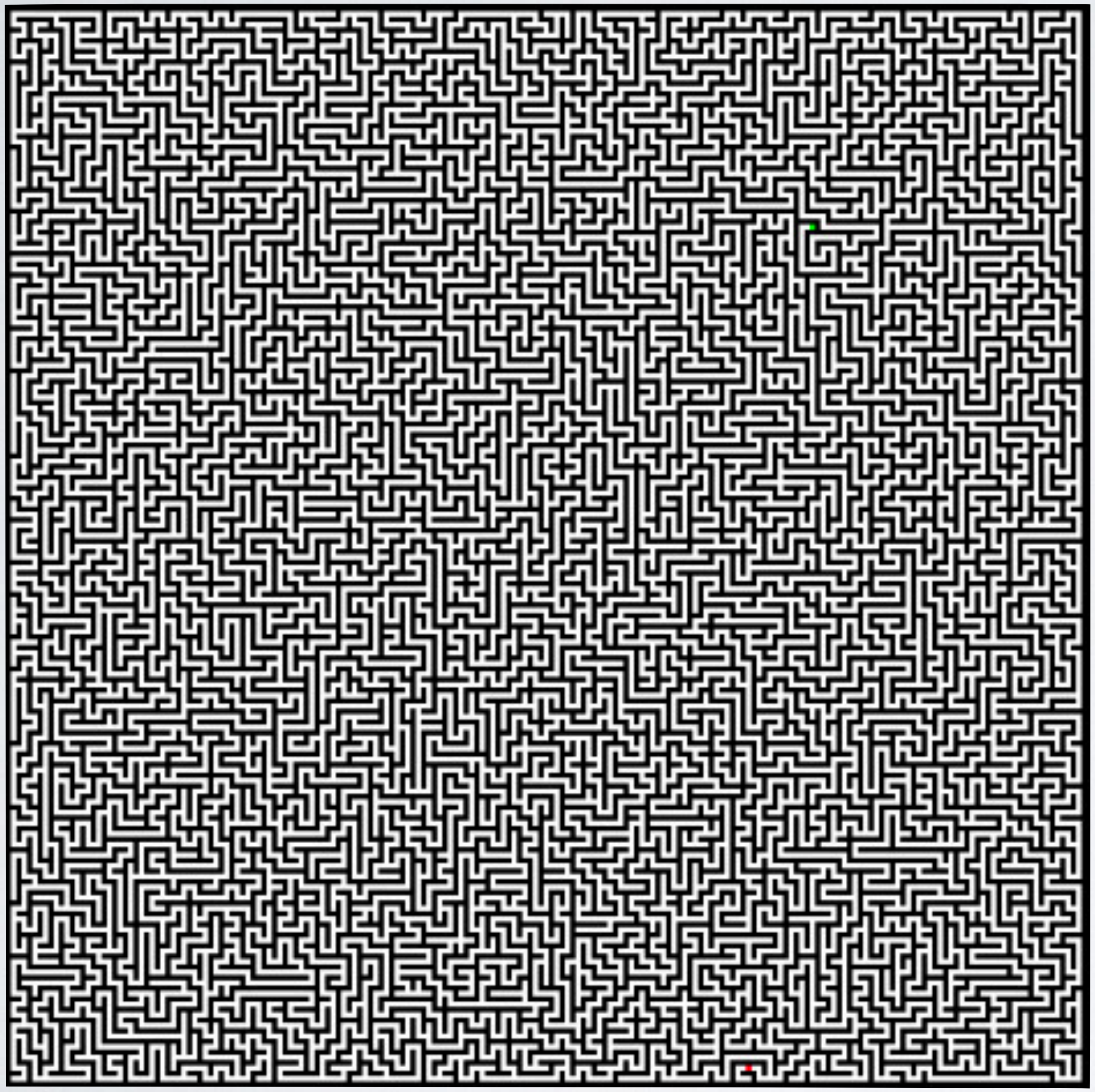


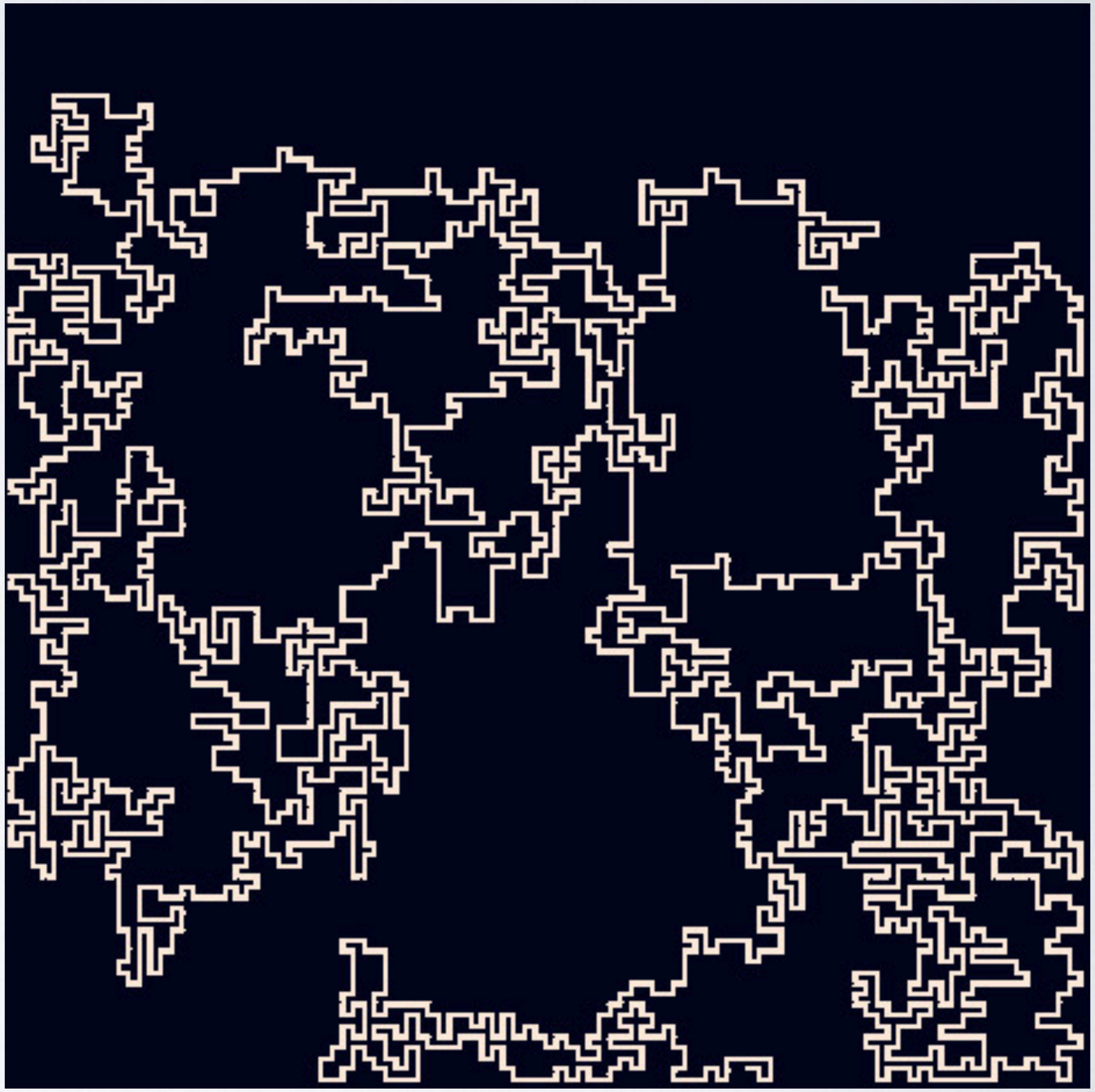
Schwarzschild et al. "Datasets for Studying Generalization from Easy to Hard Examples"

Train on 9x9 → Test on 201x201

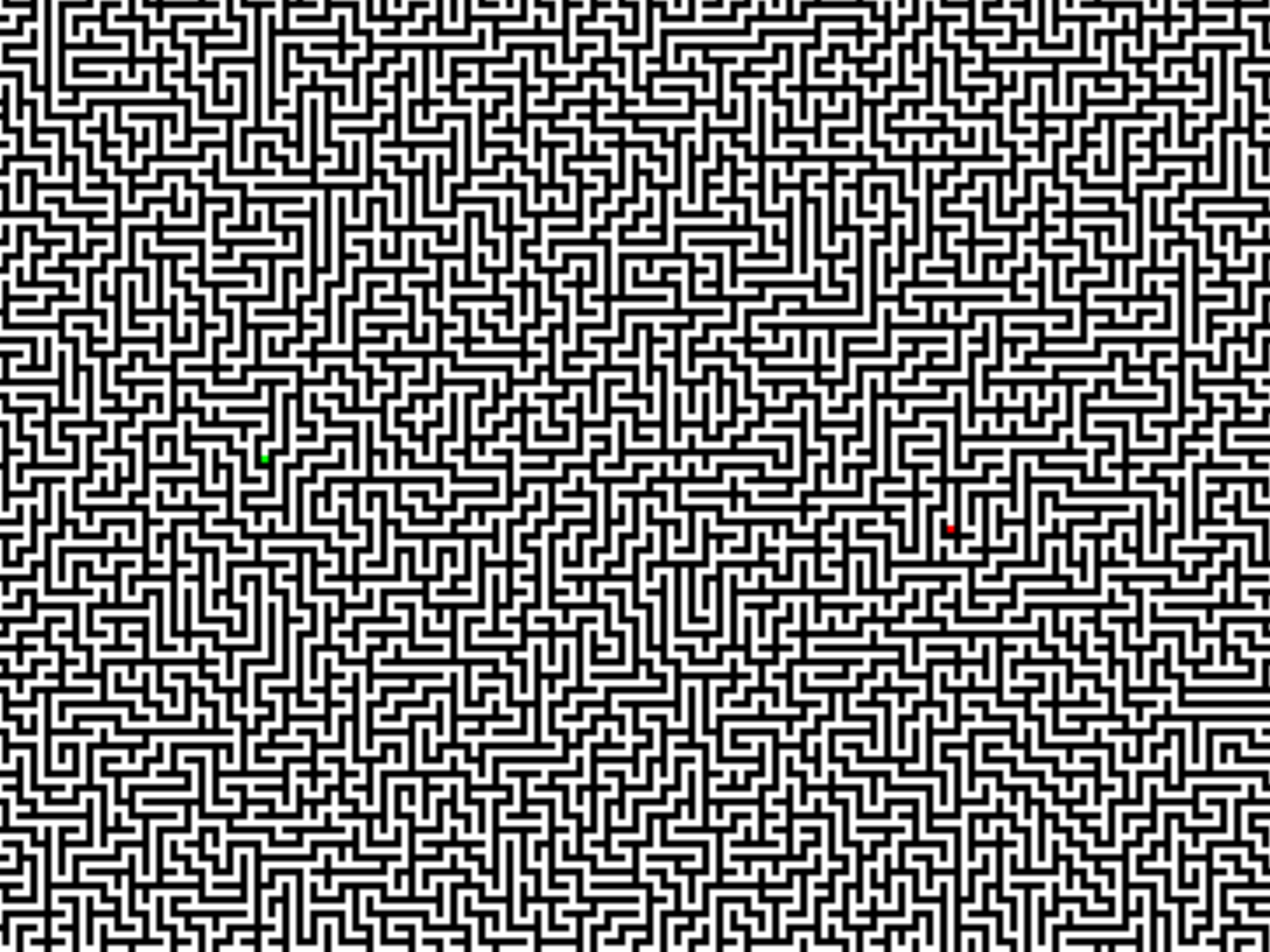
30 iterations

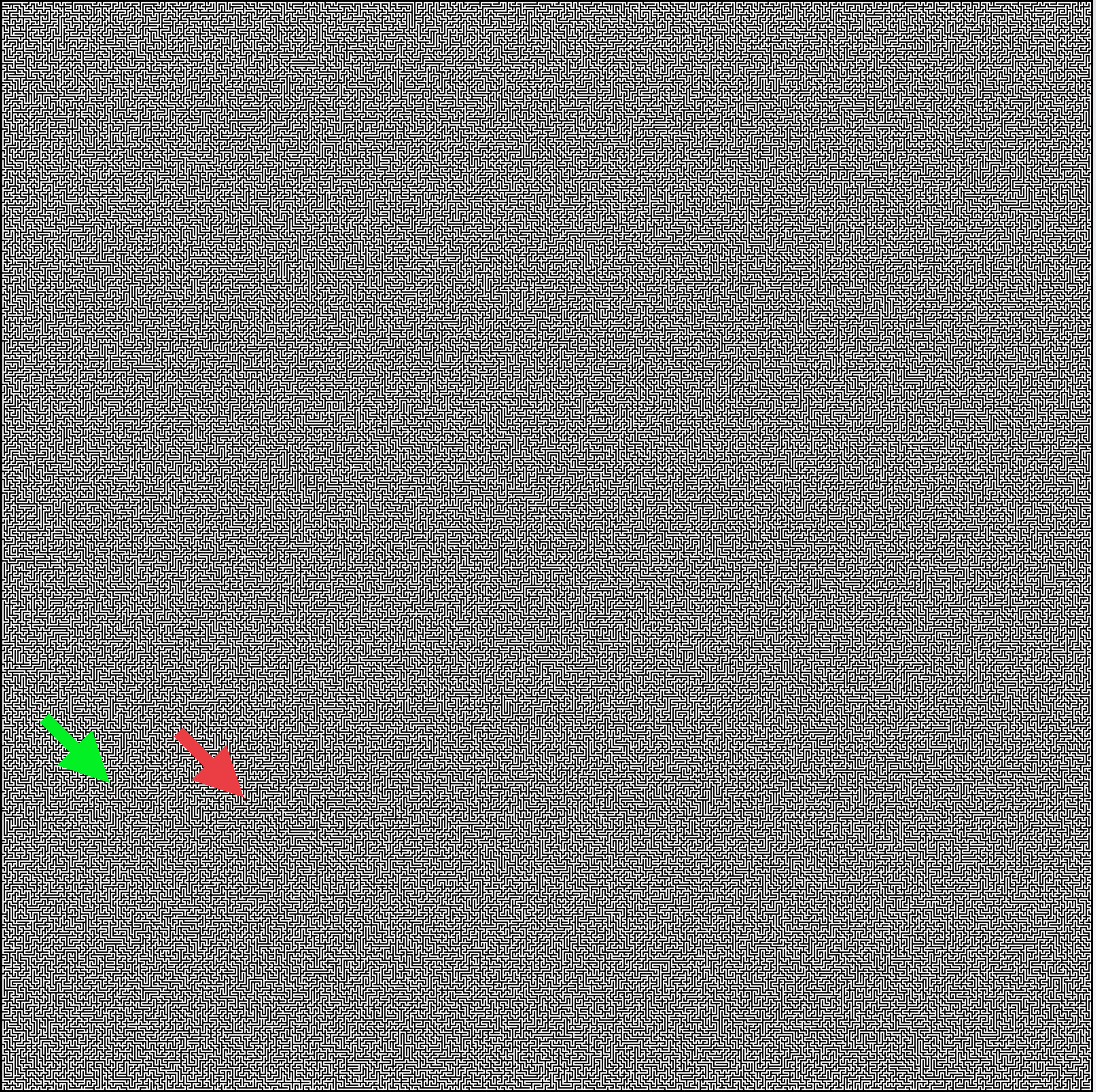
2400 iterations



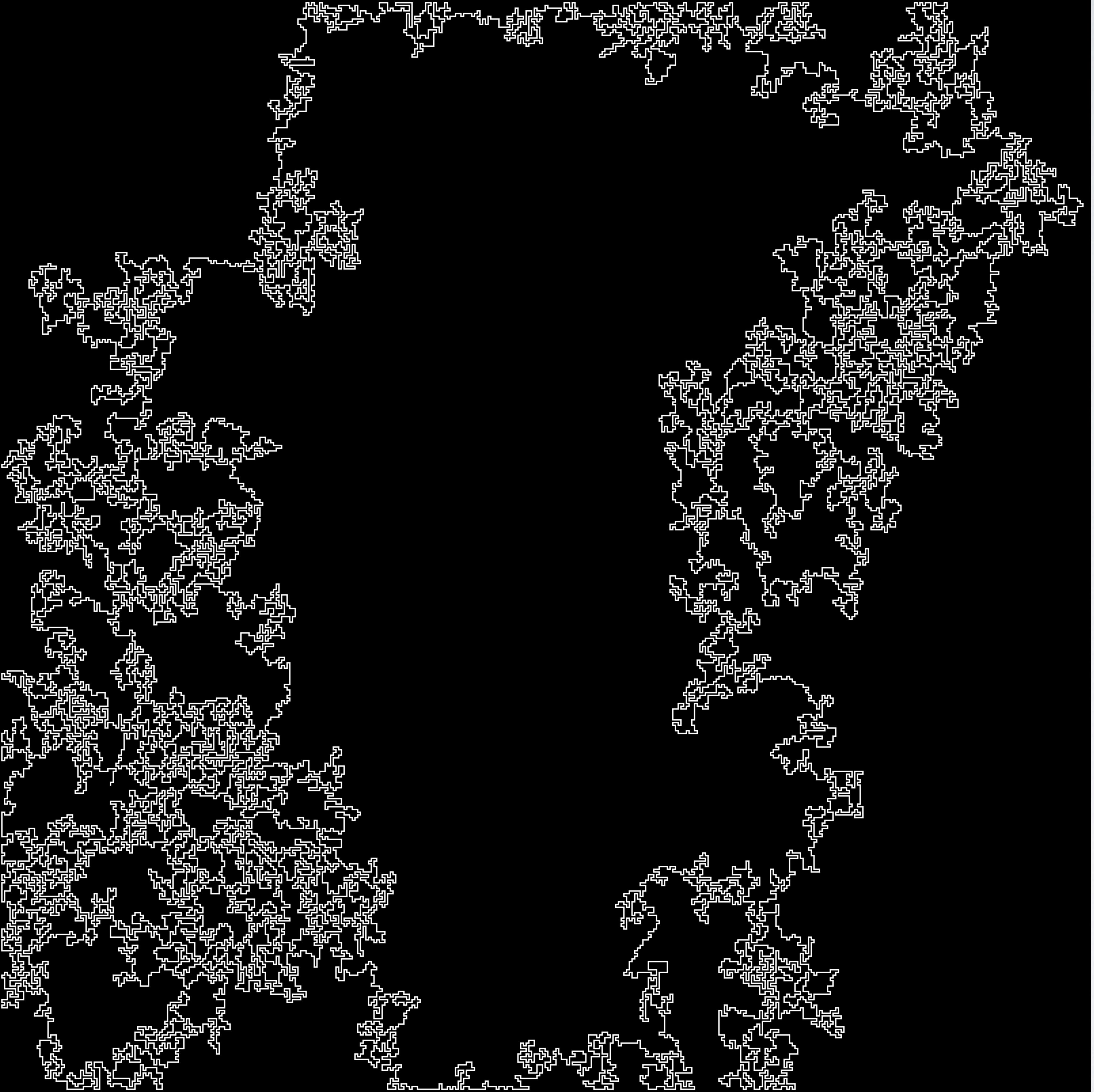


Schwarzschild et al. "Datasets for Studying Generalization from Easy to Hard Examples"





801x801



801x801

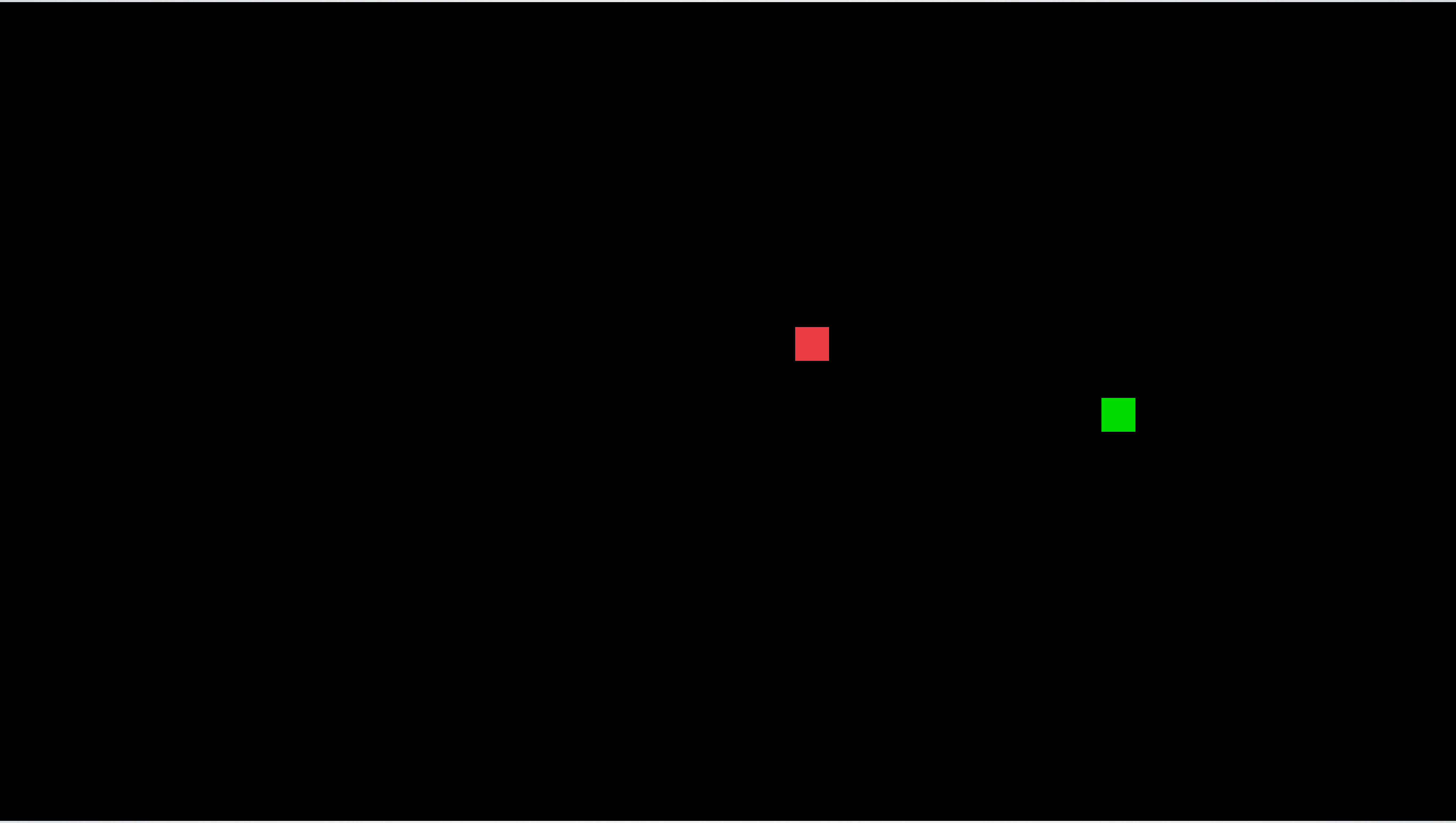
20,000 “thoughts”

100,004 layers

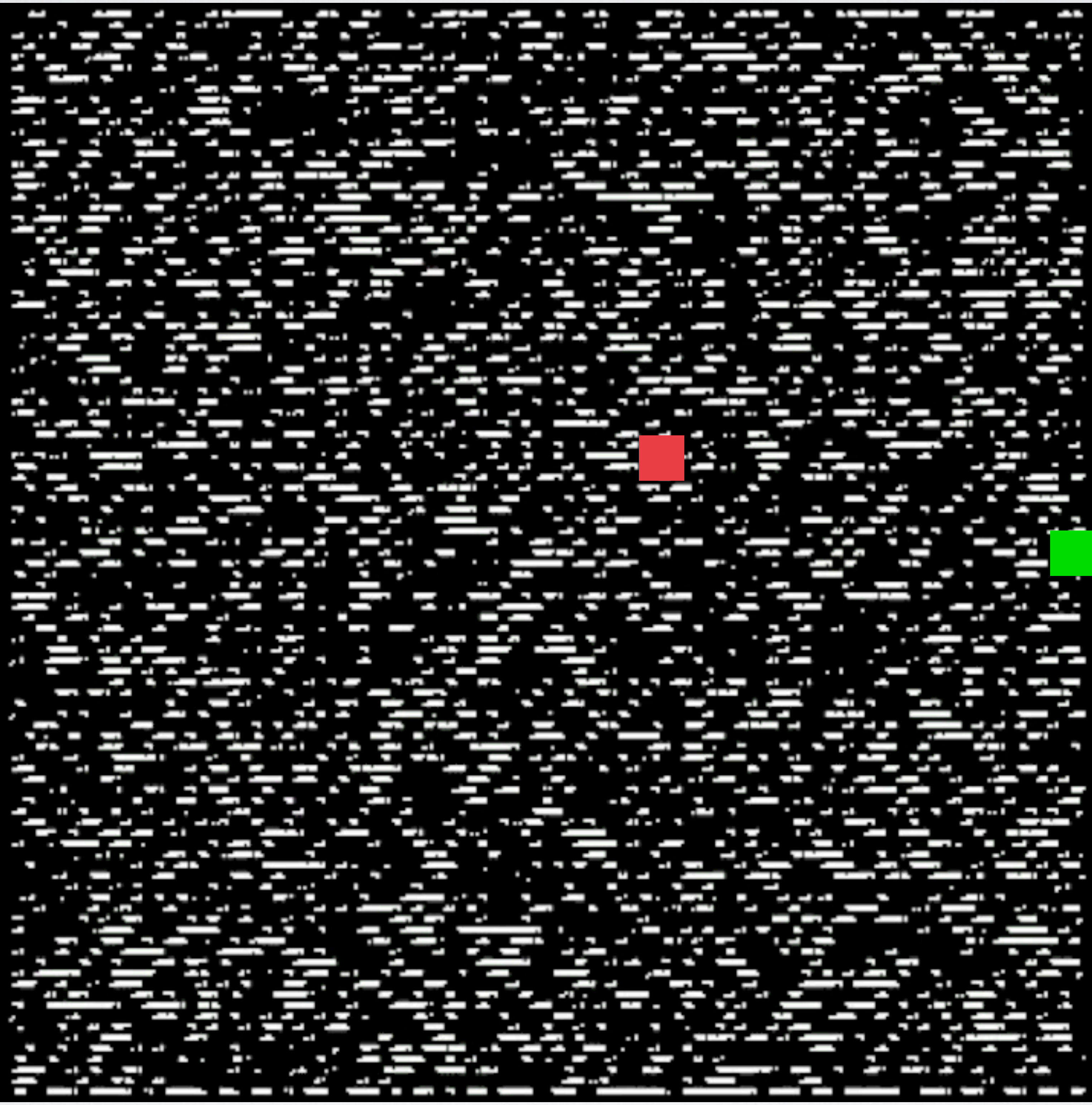
1 (trivial) pixel error

Testing the **robustness** of thinking systems

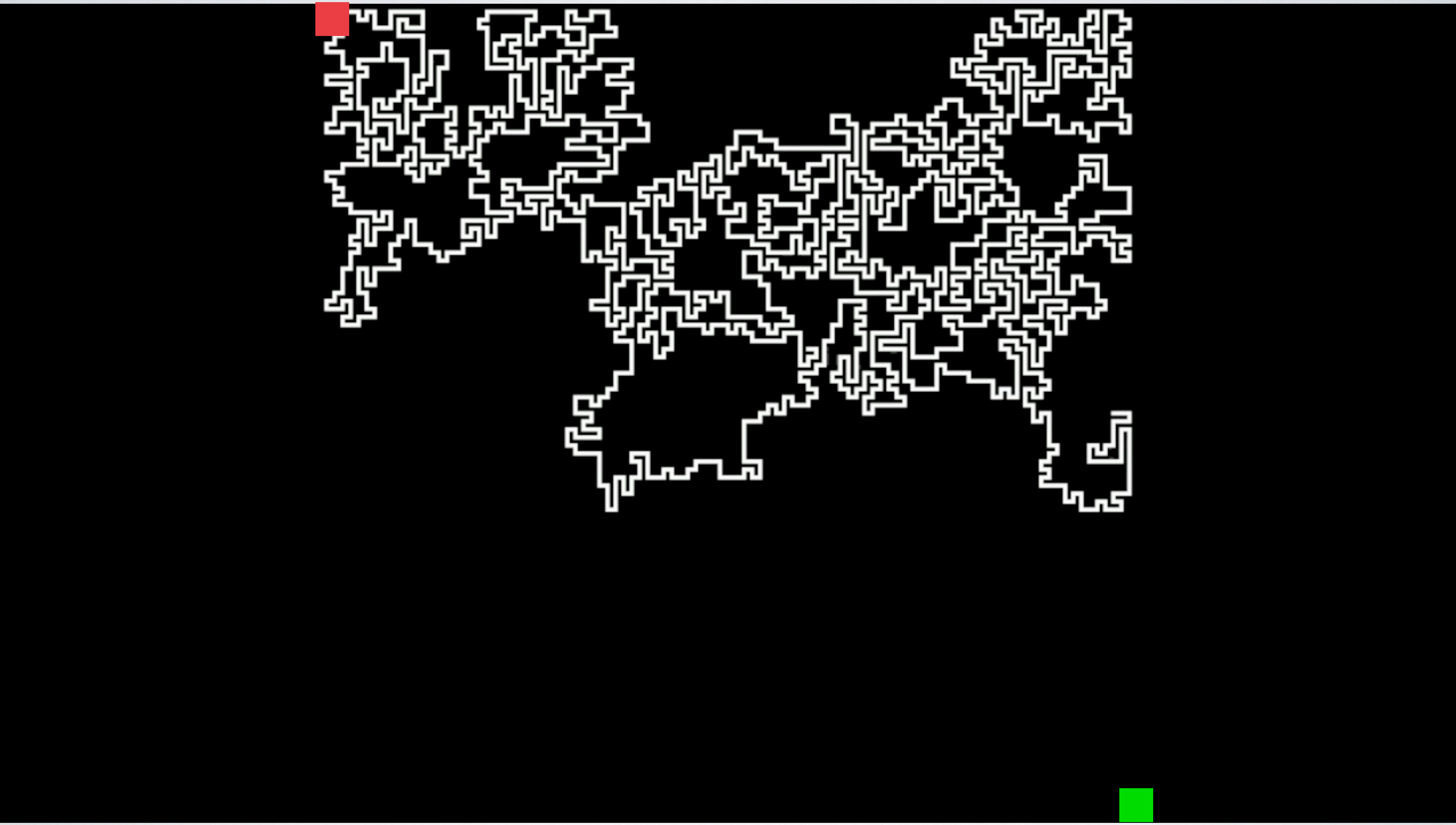
Solving a maze: start to finish



Corrupt memory with Gaussian noise



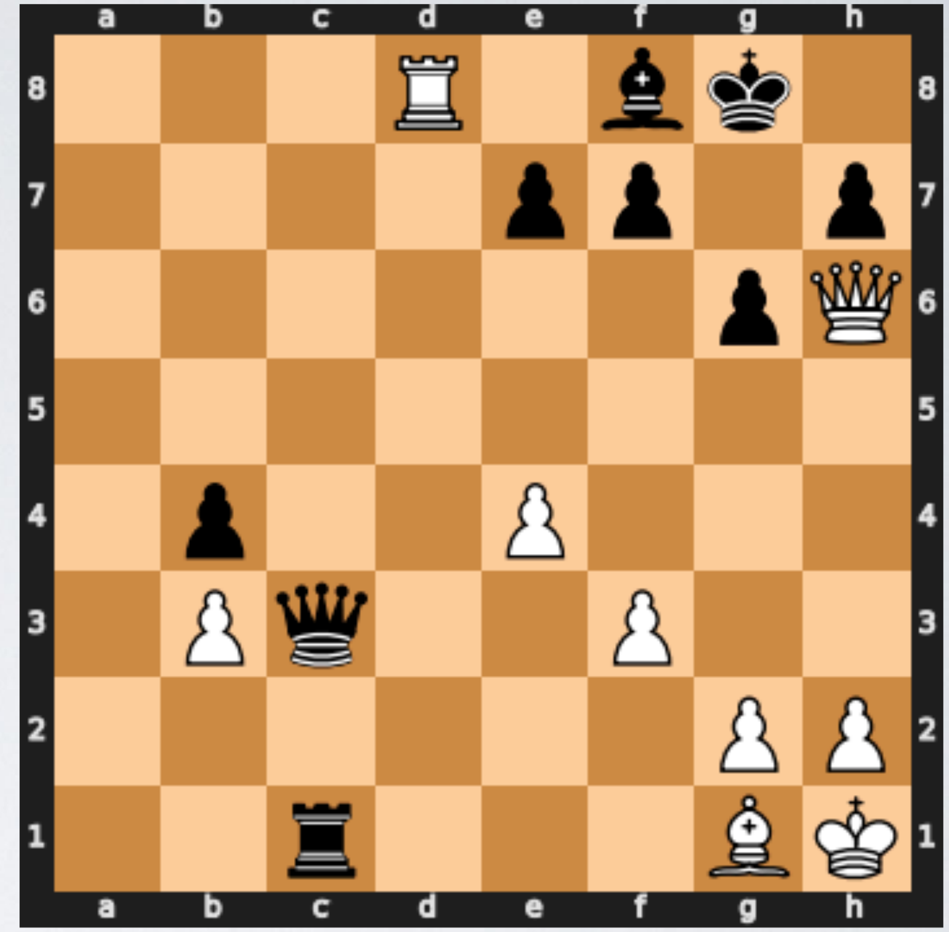
Change the maze entry and exit point



CHALLENGE PROBLEM

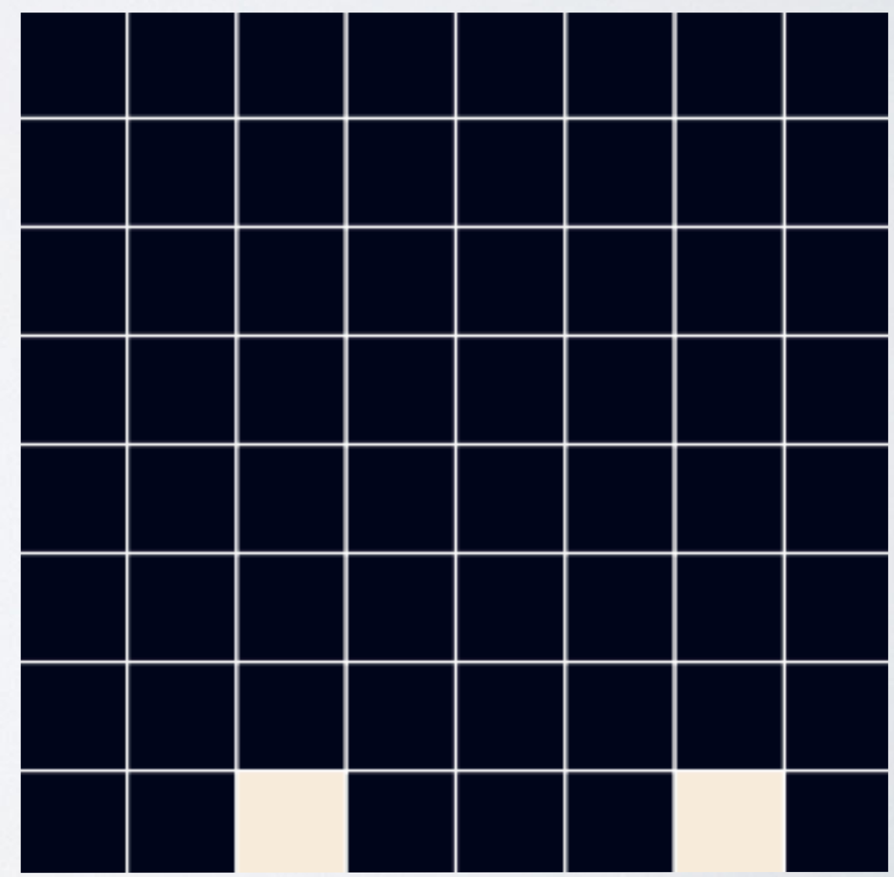
Chess

lichess.org



"Chess puzzles"

Game scenarios that have clear "best move".



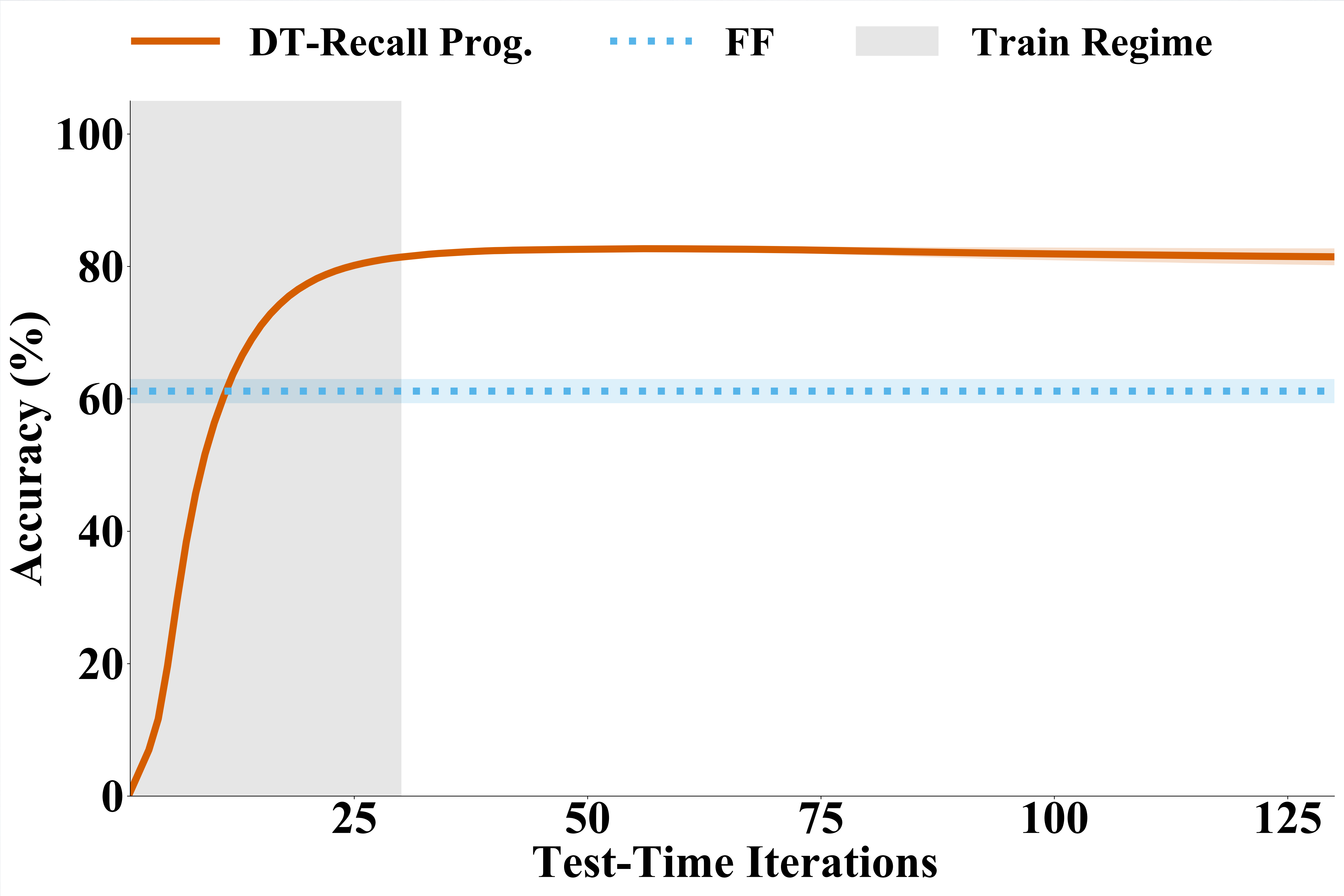
Each puzzle has an Elo rating from human play.

But what happens when they "think for longer?"

1 million puzzles

Train on “easy”, test on “hard”

Chess



Some thoughts about thinking…

Thinking systems see only the *problem* and *solution*, and organically learn algorithms end-to-end.

Thinking systems generalize to "hard" problems that lie outside the training distribution.

Thinking systems can potentially replace hand-crafted algorithms in ML systems.

Thanks!