

Benchmarks & Evaluation

6.S891/12.S992/6.S893: AI for Climate Action

Spring 2026

Speaker: Sara Beery

(Some content was originally created with Elijah Cole for the CV4Ecology program)

Machine Learning: A Success Story

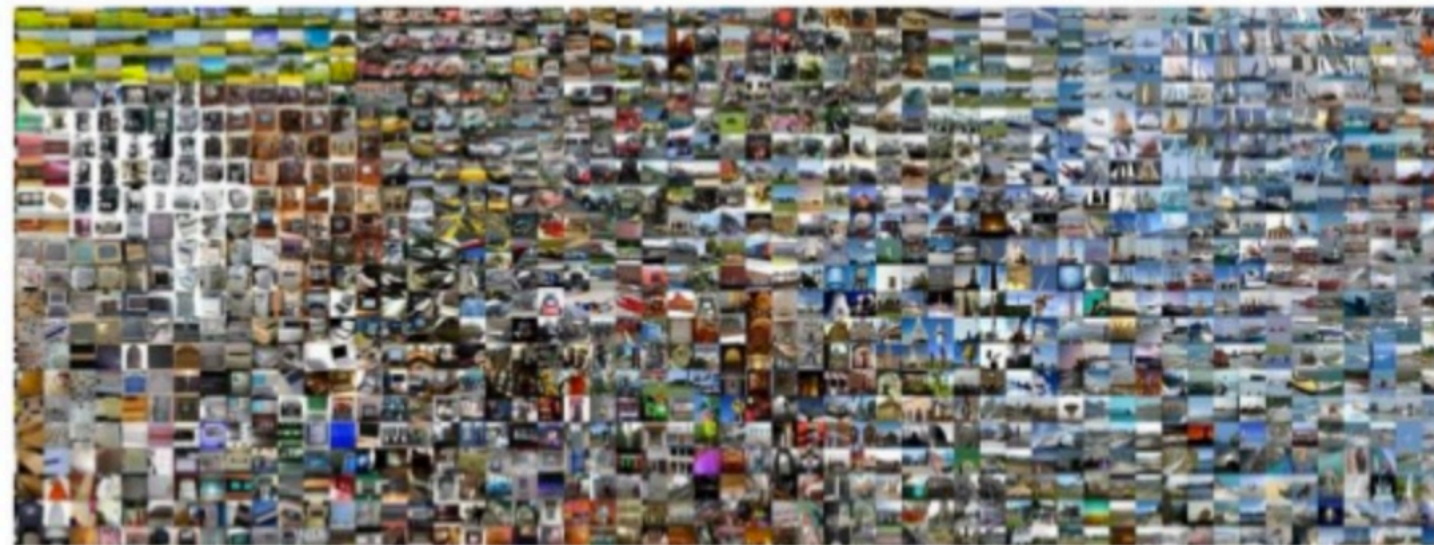


Image Classification

Input sentence:	Translation (PBMT):	Translation (GNMT):	Translation (human):
李克強此行將啟動中加總理年度對話機制，與加拿大總理杜魯多舉行兩國總理首次年度對話。	Li Keqiang premier added this line to start the annual dialogue mechanism with the Canadian Prime Minister Trudeau two prime ministers held its first annual session.	Li Keqiang will start the annual dialogue mechanism with Prime Minister Trudeau of Canada and hold the first annual dialogue between the two premiers.	Li Keqiang will initiate the annual dialogue mechanism between premiers of China and Canada during this visit, and hold the first annual dialogue with Premier Trudeau of Canada.

Machine Translation



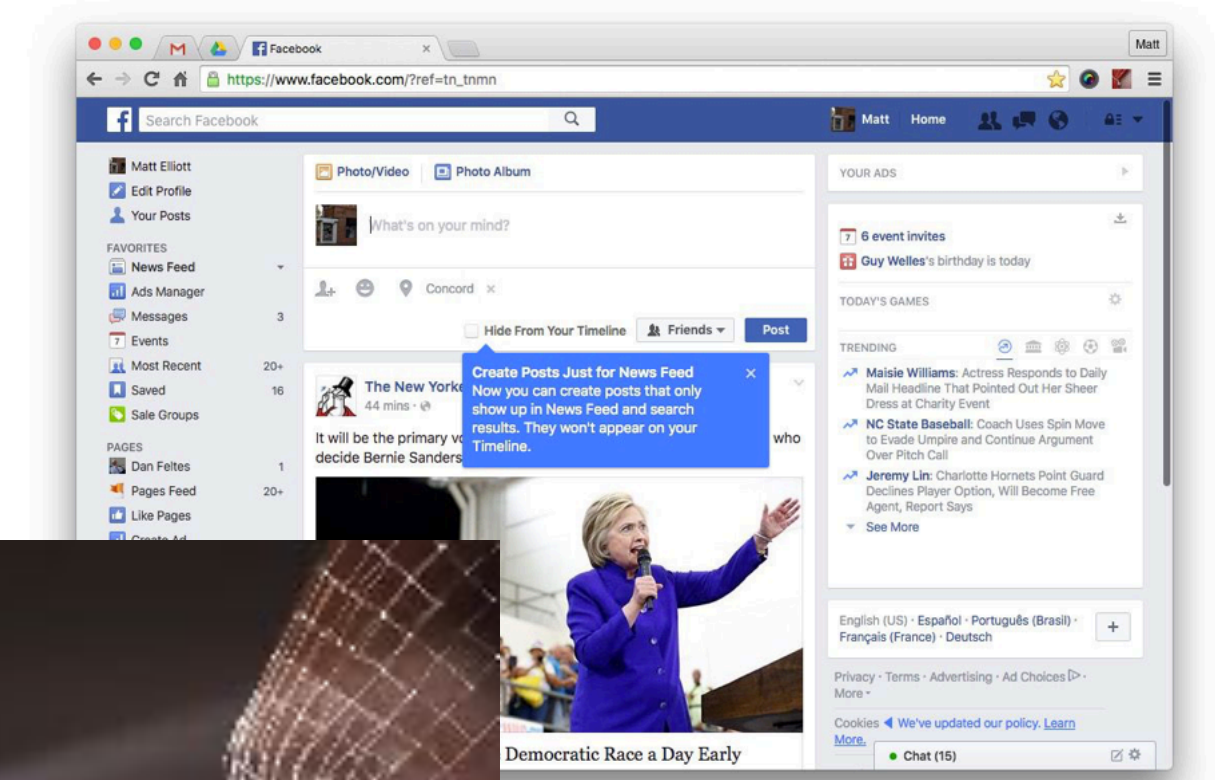
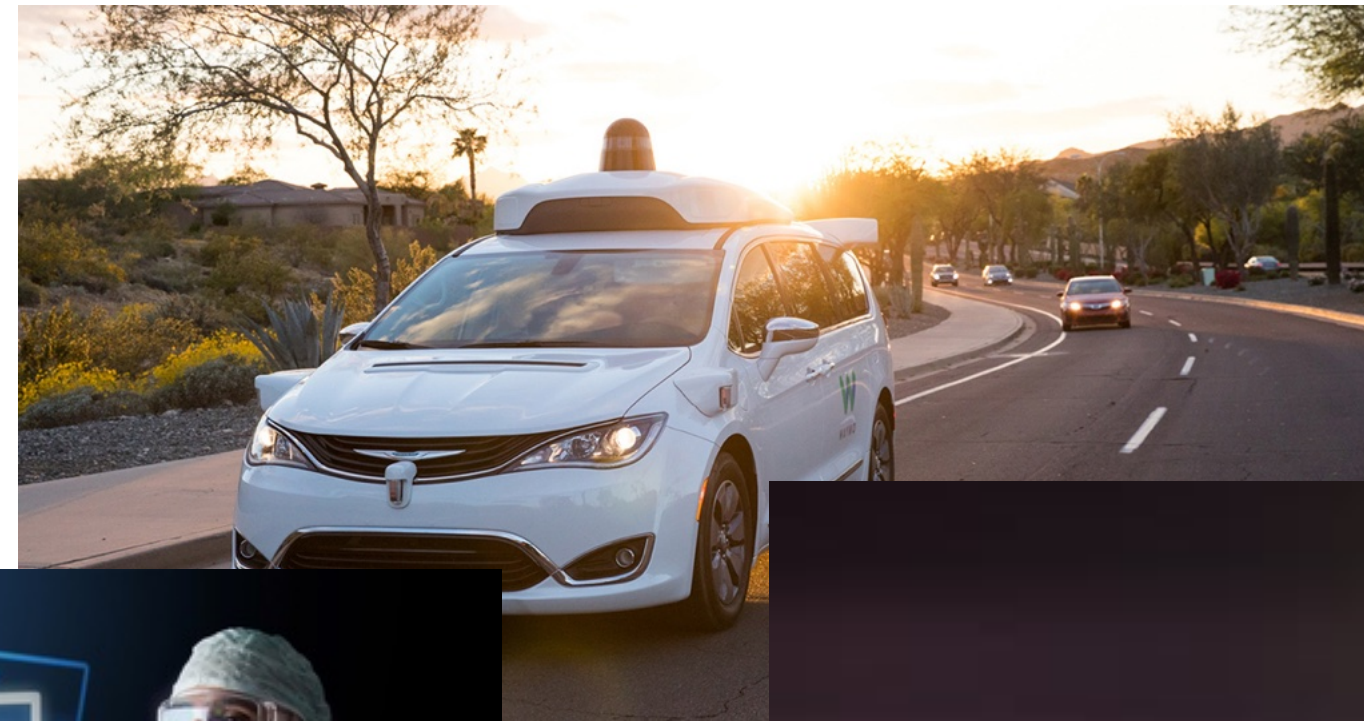
Strategy Games



Realistic Image Generation

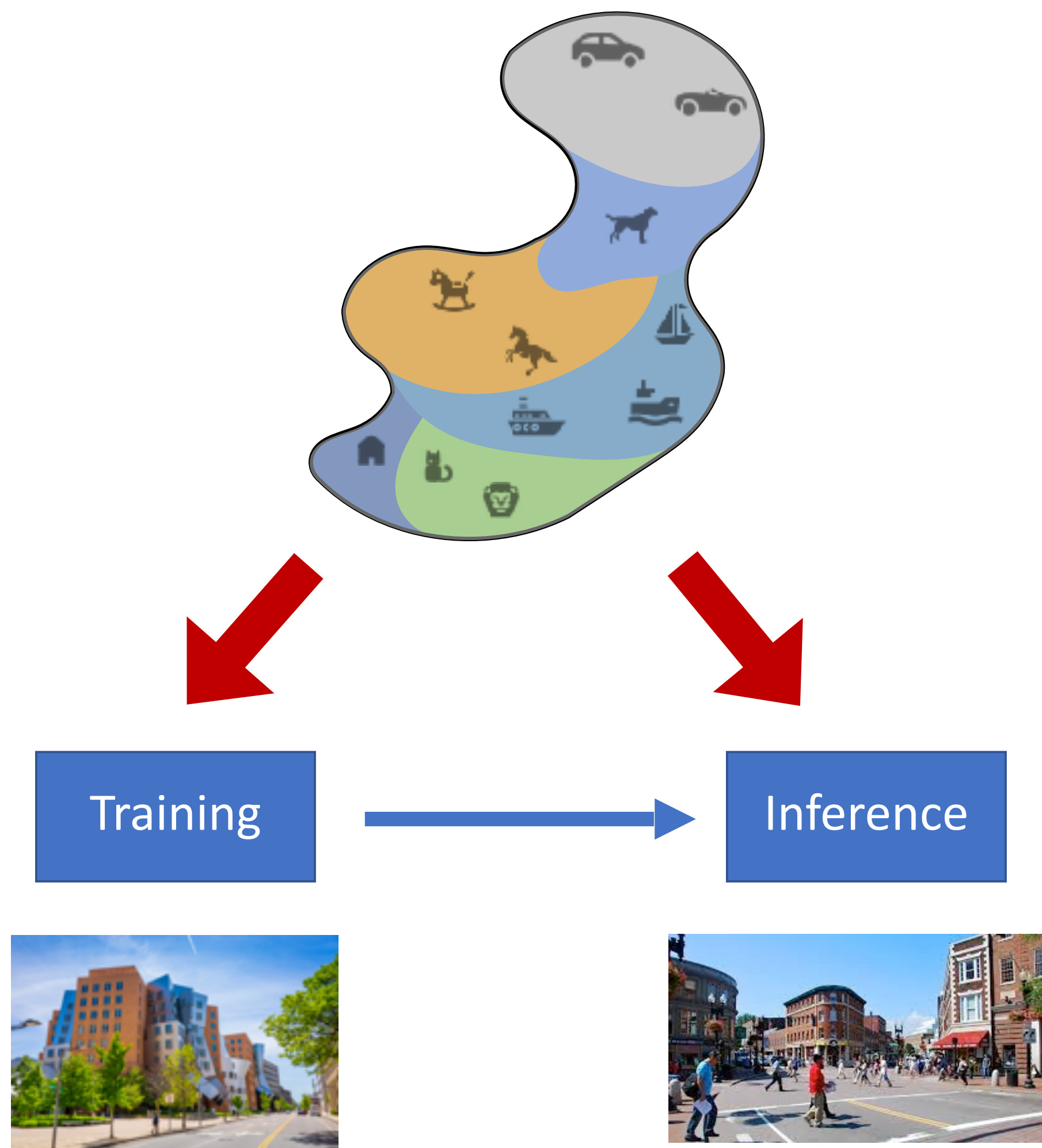


Robotic Manipulation



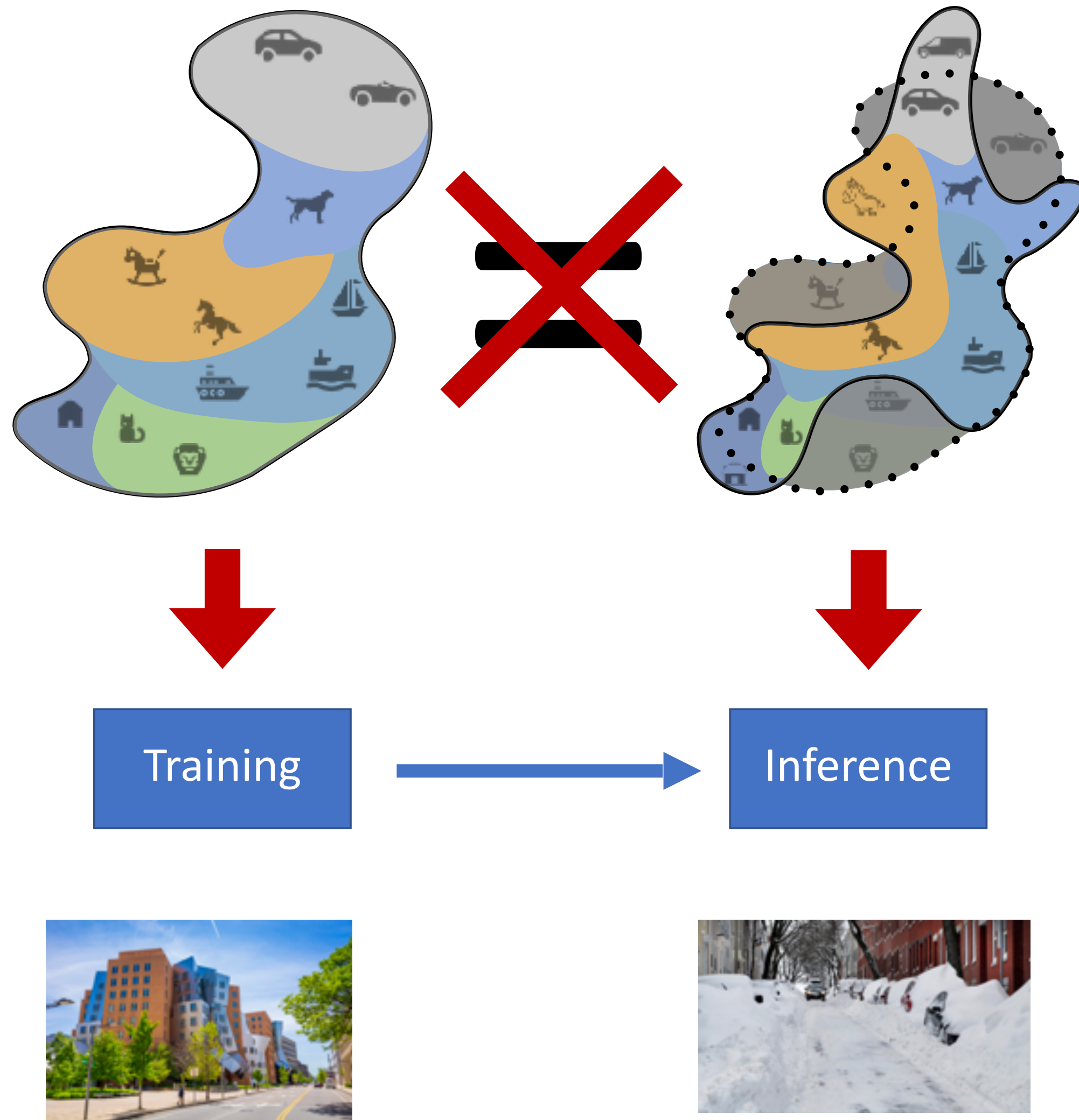
Do ML systems work in the real world?

Standard ML setting



training distribution
=
test distribution

... vs the real world



deploy model on data from a different distribution

e.g.:


- perturbed data
- different label distribution
- other shifts (sequence/graph size, weather, country/city, source of measurement,...)

What can go wrong?

hardmaru @hardmaru · Sep 9

The new Roomba uses AI to avoid smearing dog poop all over your house.

"But in order to make this possible, the company first had to create a diverse dataset of poop."



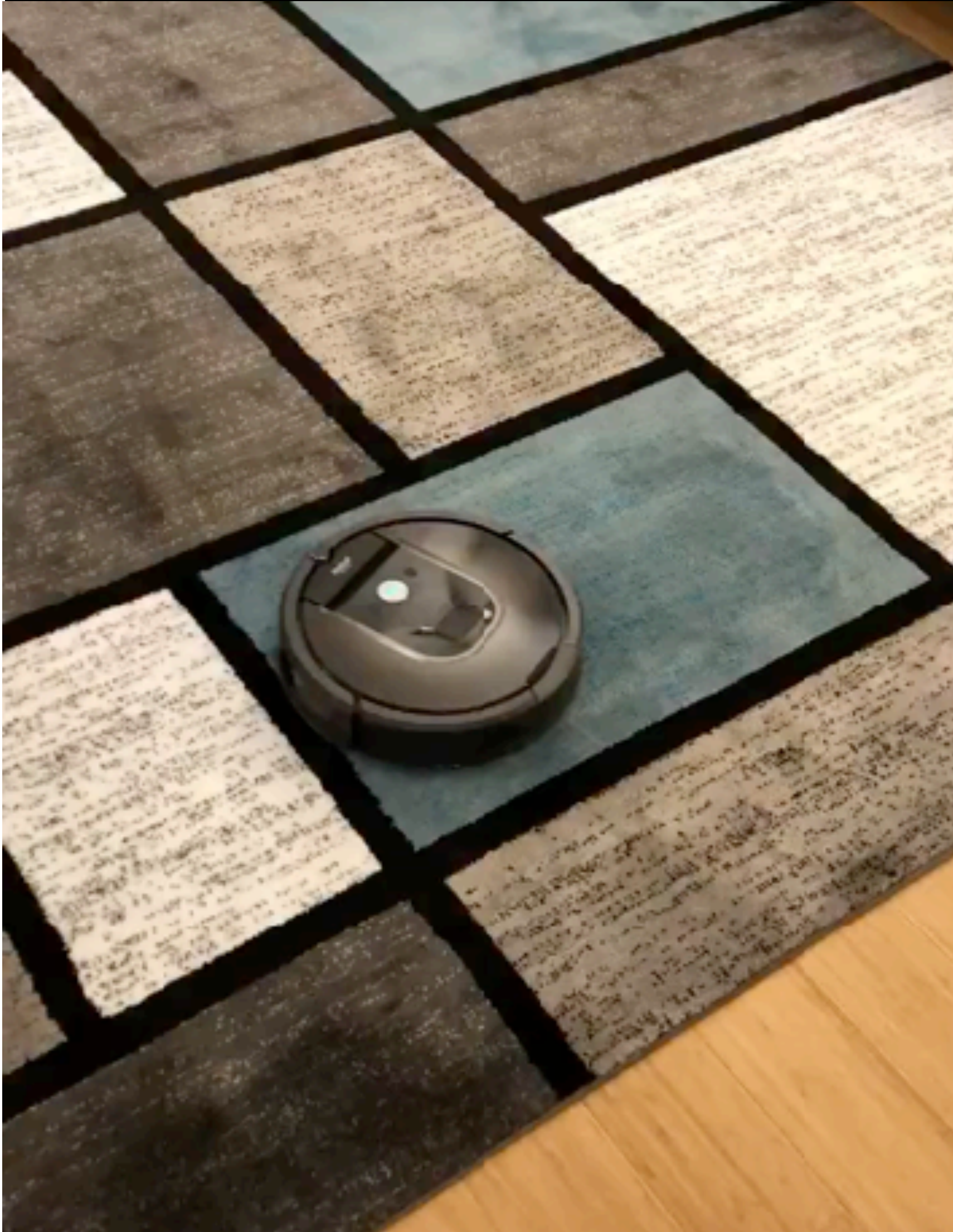
The new Roomba uses AI to avoid smearing dog poop all over your ho...
iRobot, the company that makes the Roomba, is trying to eliminate messy accidents with the use of artificial intelligence. On Thursday ...
[edition.cnn.com](#)

62 336 1.3K

Dmitry Krotov
@DimaKrotov

Replying to @hardmaru

I wish they had also created a diverse dataset of rugs so that it didn't confuse black stripes with cliffs and I could finally get my entire house cleaned 😂



Concrete Problems in AI Safety

Dario Amodei*
Google Brain

Chris Olah*
Google Brain

Jacob Steinhardt
Stanford University

Paul Christiano
UC Berkeley

John Schulman
OpenAI

Dan Mané
Google Brain

might serve a benchmarking role similar to that of the bAbI tasks [163], with the eventual goal being to develop a single architecture that can learn to avoid catastrophes in all environments in the suite.

7 Robustness to Distributional Change

All of us occasionally find ourselves in situations that our previous experience has not adequately prepared us to deal with—for instance, flying an airplane, traveling to a country whose culture is very different from ours, or taking care of children for the first time. Such situations are inherently



Center for Statistics
and Machine Learning



A Princeton University Workshop

The Reproducibility Crisis In ML-based Science

28th
July 2022
Online

Session 1: 10 AM - 12 PM ET

DIAGNOSE

Dr. Michael Roberts
University of Cambridge

Dr. Gilles Vandewiele
Ghent University

Prof. Odd Erik Gundersen
NTNU

Session 2: 12 PM - 2 PM ET

FIX

Prof. Michael Lones
Heriot-Watt University

Prof. Marta Serra-Garcia
UC San Diego

Dr. Momin Malik
Mayo Clinic

Session 3: 2 PM - 4 PM ET

FUTURE PATHS

Dr. Jake Hofman
Microsoft Research

Prof. Brandon Stewart
Princeton University

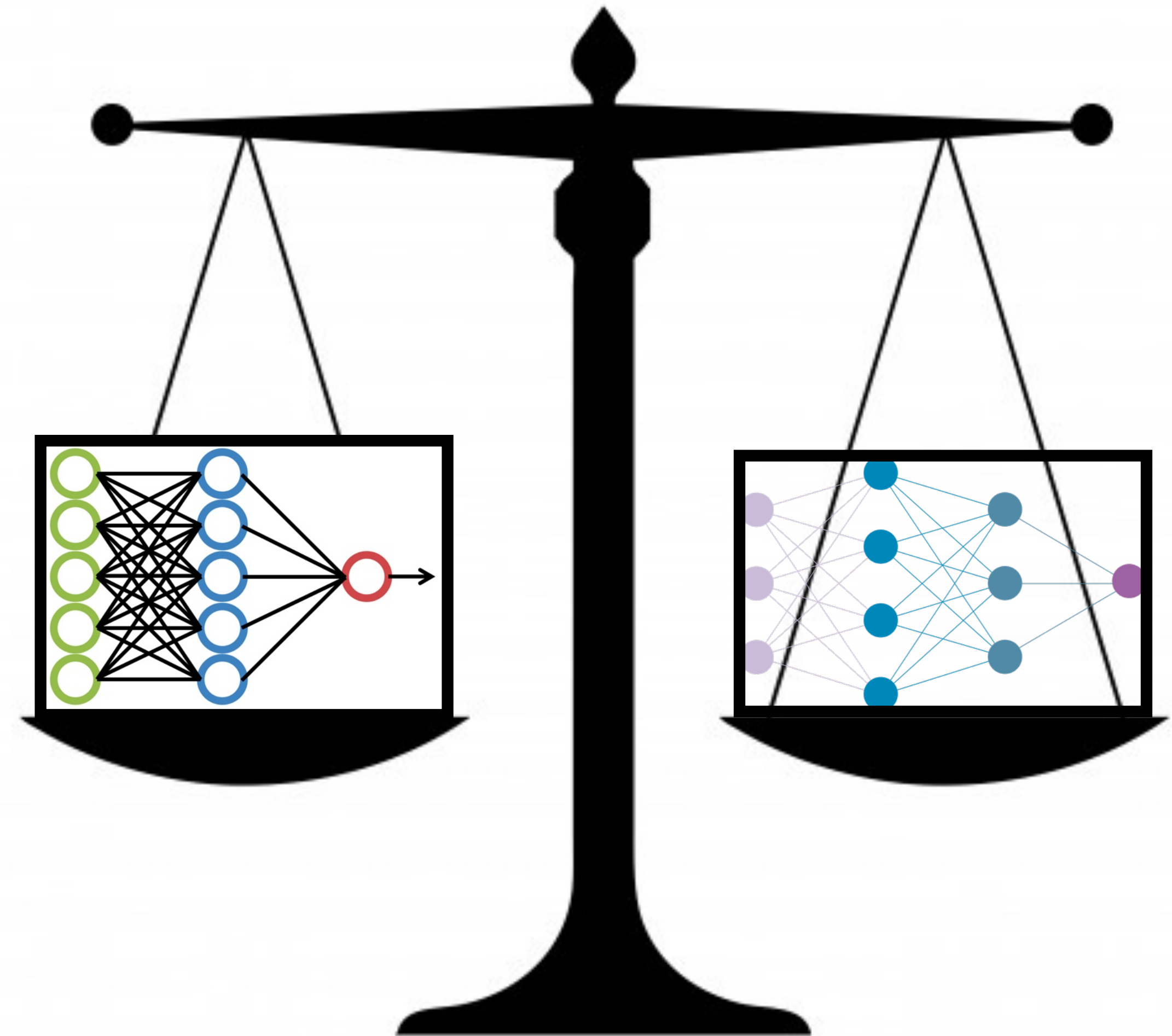
Prof. Jessica Hullman
Northwestern University

Details & RSVP at
bit.ly/rep-workshop

Organizers: Sayash Kapoor Priyanka Nanayakkara Kenny Peng Hien Pham Arvind Narayanan

Benchmarking & Evaluation

- Benchmarking
- Metrics
- Fair comparisons
- Ablations
- Evaluating generative models
- Saturated benchmarks?
- Where do models fail?



Benchmarking



Benchmarks have played an integral role in deep learning research



airplane

automobile

bird

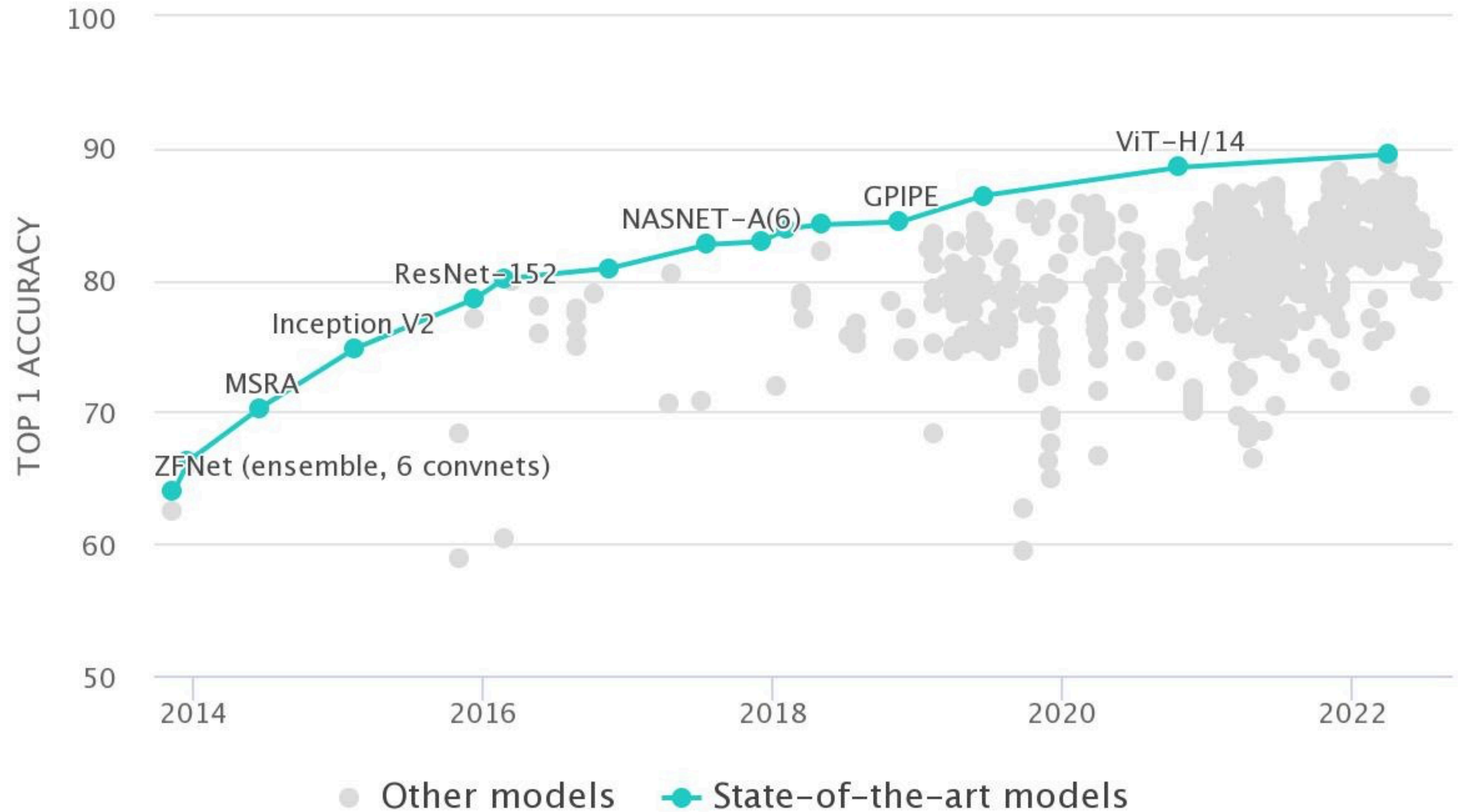
cat

deer

dog



They help us measure progress in the field





They are a fundamental aspect of formalizing new research questions and open challenges



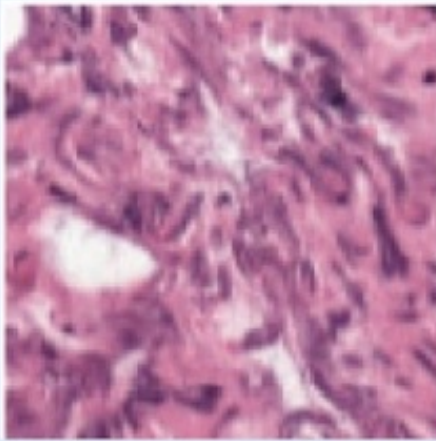



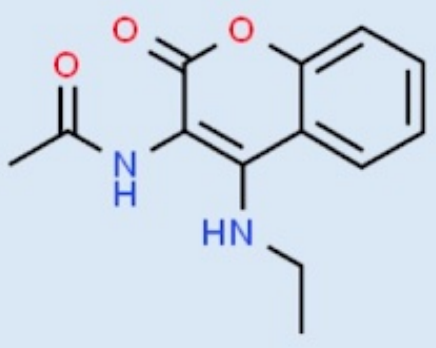
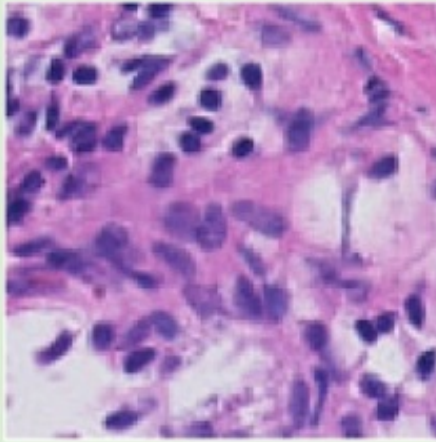



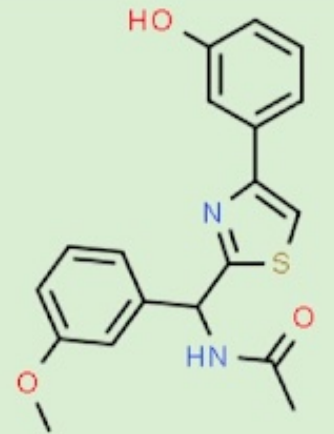
What makes up a benchmark?

- A dataset (at least one)
- A split of that data into train, validation, test (at least one)
- A metric to optimize (at least one)

The structure of a benchmark defines the challenge

WILDS

Pang Wei Koh*, Shiori Sagawa*, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanus Phillips, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang

	Camelyon17	iWildCam	PovertyMap	FMoW	Amazon	CivilComments	OGB-MolPCBA
Shift	Hospitals	Locations	Countries	Time	Users	Demographics	Scaffold
Train					Overall a solid package that has a good quality of construction for the price.	What do Black and LGBT people have to do with bicycle licensing?	
Test					I *loved* my French press, it's so perfect and came with all this fun stuff!	As a Christian, I will not be patronizing any of those businesses.	
Adapted from	Bandi et al. 2018	Beery et al. 2020	Yeh et al. 2020	Christie et al. 2018	Ni et al. 2019	Borkan et al. 2019	Hu et al. 2020

The structure of a benchmark defines the challenge

Training data

Camera 1



Camera 2



...

Camera 245



Out-of-distribution (OOD) test data

Camera 246



...

Control: In-distribution (ID) test data

Camera 1

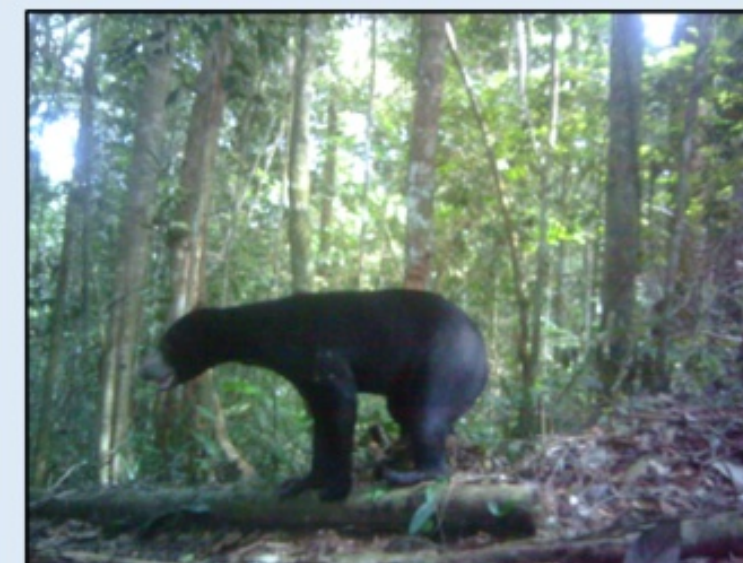


Camera 2



...

Camera 245



The structure of a benchmark defines the challenge

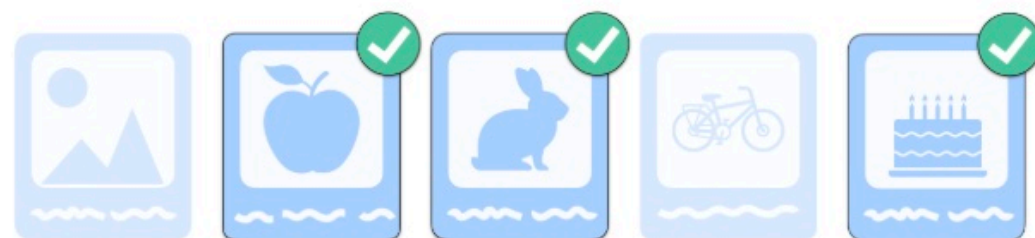


Welcome to DataComp, the machine learning benchmark where the models are fixed and the challenge is to find the best possible data! Select a setting to learn more about how to participate

CLIP

Contrastive Language Image Pre-training

Select the best subset of **image/text pairs** from a large pool to train a **CLIP model**. Evaluate your training set by testing the model on a set of downstream **vision tasks**



LM

Language Modeling

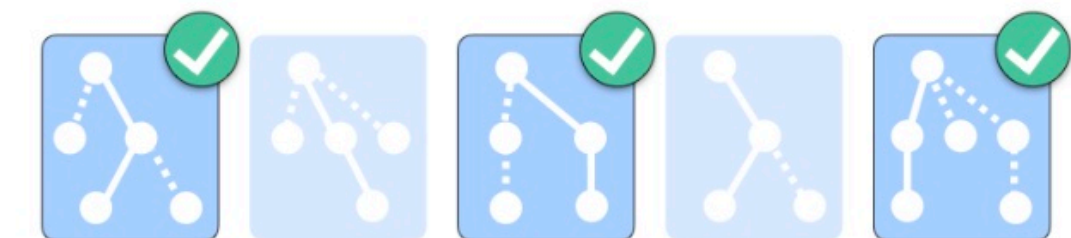
Select the best subset of **text data** from a large pool to train a **language model**. Evaluate your training set by testing the model on a set of downstream **language tasks**



Reasoning

Chain of Thought and Reasoning

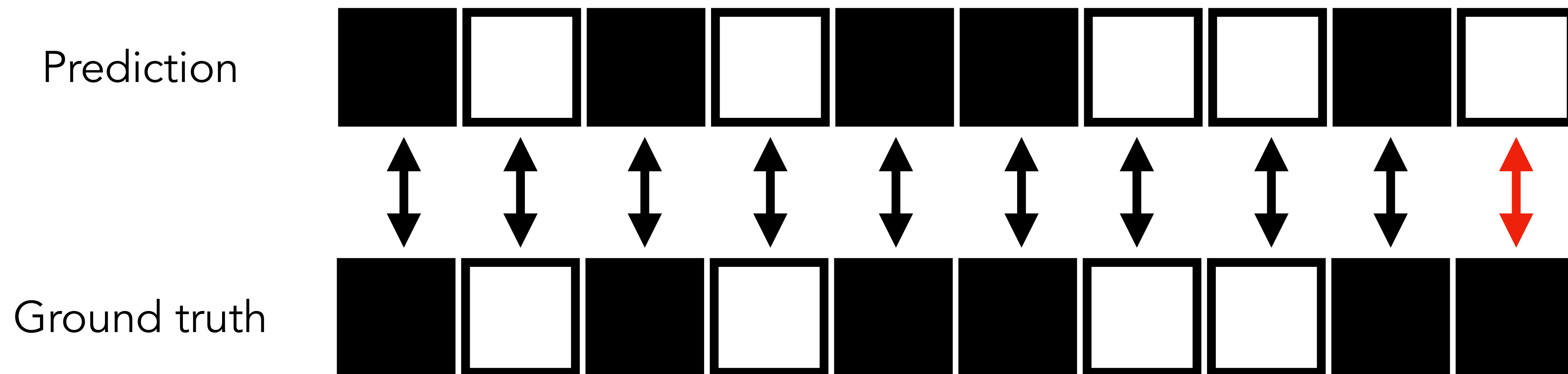
Generate the best **question - answer pairs** to teach base models to **reason**. Evaluate these models by testing them on a set of downstream **reasoning domains**



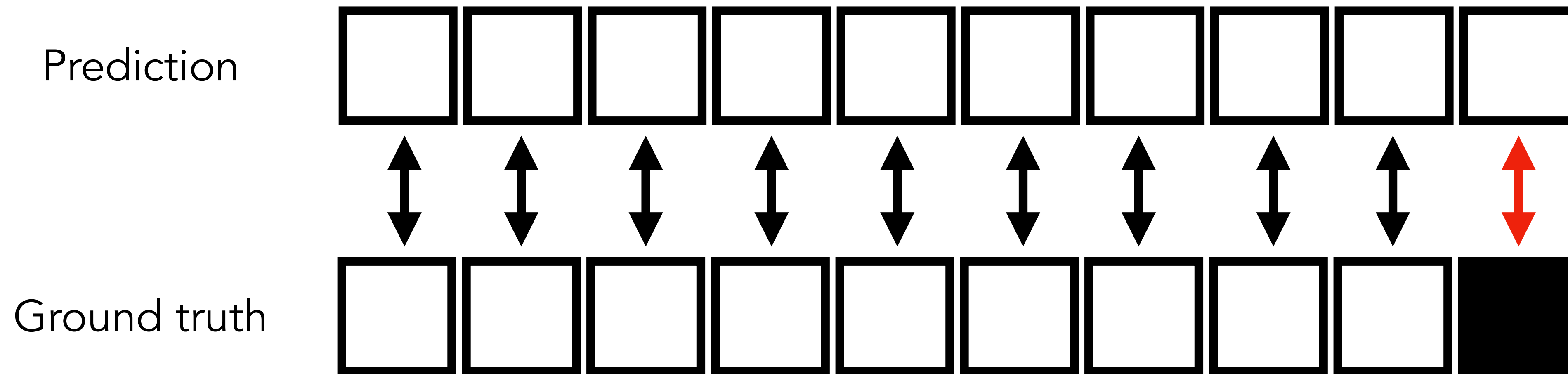
Metrics

Choosing appropriate metrics is vital

90% accuracy!

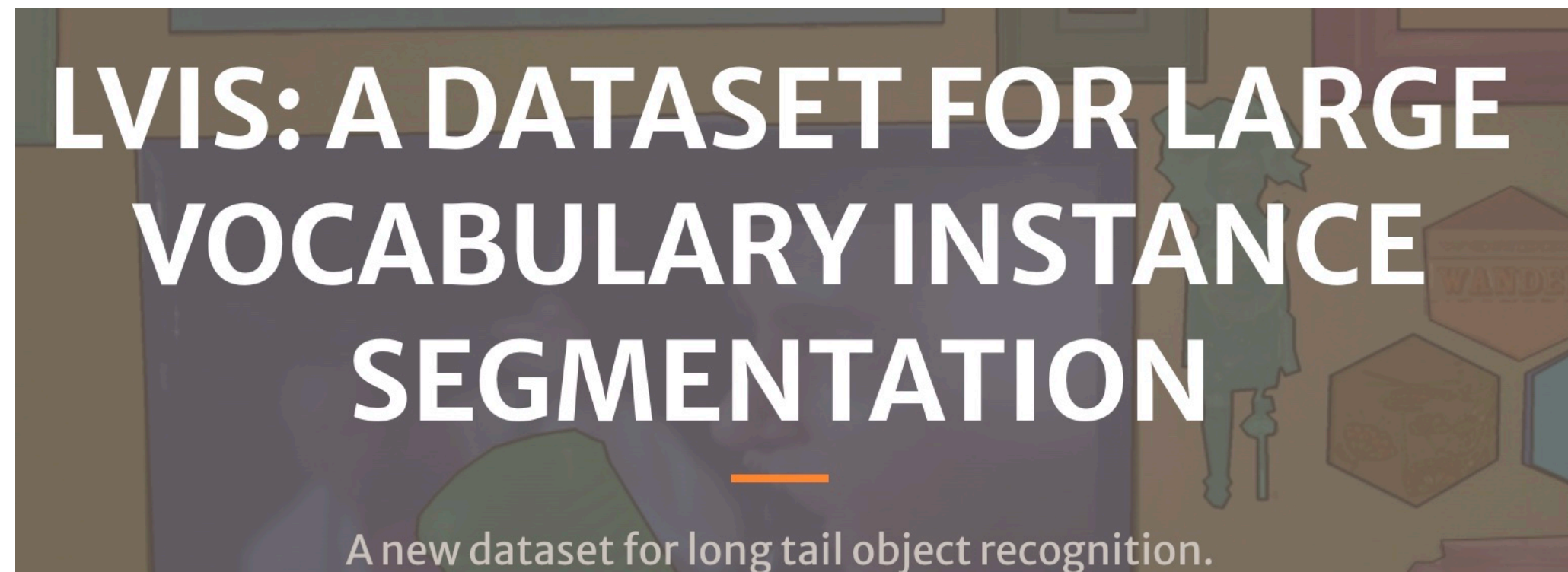


90% accuracy!



When data is imbalanced, we need metrics that account for imbalance, e.g. class average accuracy

Other ways to account for imbalance



score thr	det/img	AP	AP _r	AP _c	AP _f
0.050	100	14.8	0.8	10.9	25.3
0.050	300	15.7	0.8	12.1	26.1
0.001	300	20.8	3.3	20.7	27.9
0.000	300	20.9	3.4	20.9	27.9
0.050	100	14.8 \pm 0.19	0.6 \pm 0.21	11.0 \pm 0.36	25.2 \pm 0.10
0.000	300	21.0\pm0.17	3.2\pm0.35	21.3\pm0.45	27.7\pm0.12

Performance is reported on aggregations of "rare", "common", and "frequent" categories

<https://www.lvisdataset.org/>



2. Metrics

The following 12 metrics are used for characterizing the performance of an object detector on COCO:

Average Precision (AP):

- AP % AP at IoU=.50:.05:.95 (**primary challenge metric**)
- AP^{IoU=.50} % AP at IoU=.50 (PASCAL VOC metric)
- AP^{IoU=.75} % AP at IoU=.75 (strict metric)

AP Across Scales:

- AP^{small} % AP for small objects: area < 32²
- AP^{medium} % AP for medium objects: 32² < area < 96²
- AP^{large} % AP for large objects: area > 96²

Average Recall (AR):

- AR^{max=1} % AR given 1 detection per image
- AR^{max=10} % AR given 10 detections per image
- AR^{max=100} % AR given 100 detections per image

AR Across Scales:

- AR^{small} % AR for small objects: area < 32²
- AR^{medium} % AR for medium objects: 32² < area < 96²
- AR^{large} % AR for large objects: area > 96²

<https://cocodataset.org/>

Metrics for complex tasks can be challenging to design and interpret

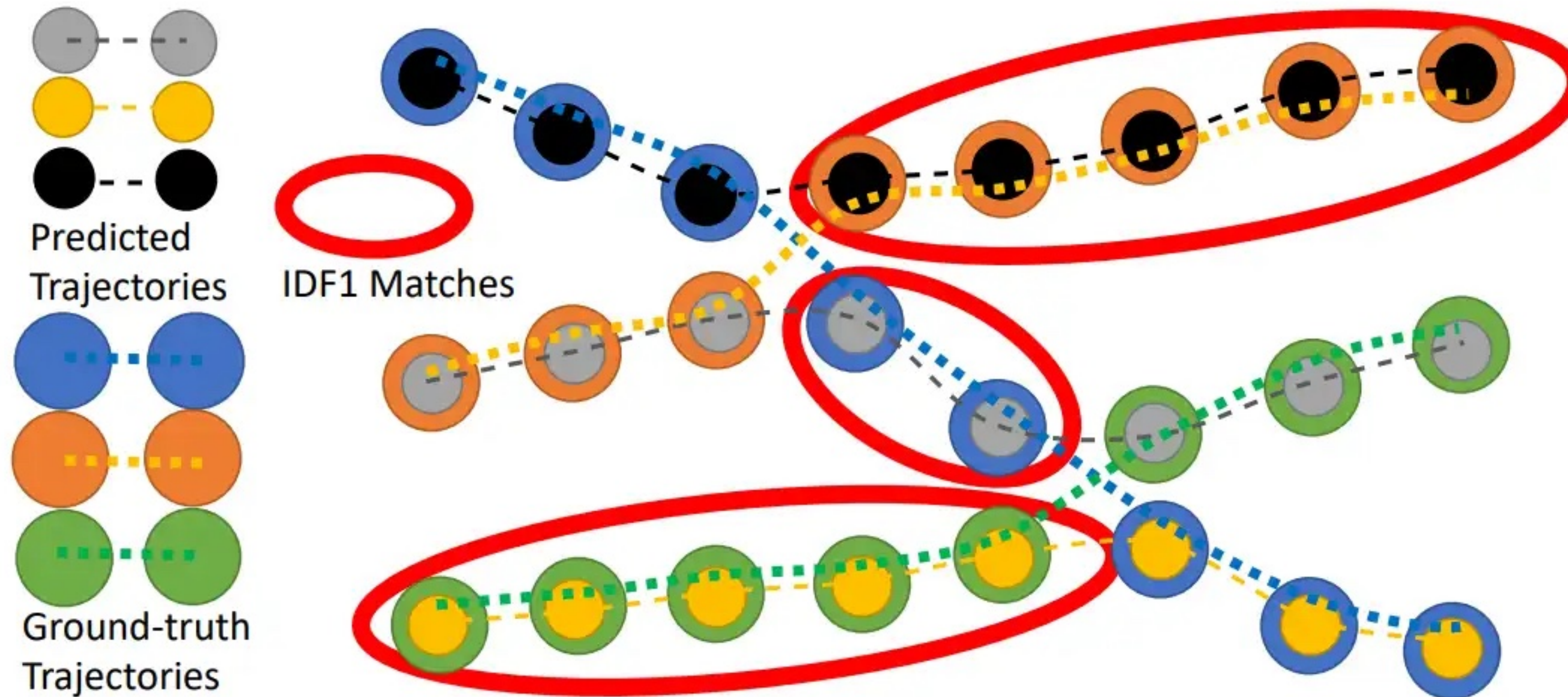


Fig. 10 A tracking example which shows how the single best trajectory matching, as performed by IDF1, can result in unintuitive matches between trajectories.

What else might we want to measure?

- Computational cost
- Speed
- Memory cost
- Training cost
- Human effort (for active/in context learning, RLHF)
- What else?

Fair comparisons

“A **fair comparison** provides evidence that *Method A* is better than *Method B* for some task.”

“A **fair comparison** provides evidence that *Method A* is better than *Method B* for some task.”

Preprocessing

Architecture

Loss

“A **fair comparison** provides evidence that *Method A* is better than *Method B* for some task.”

Preprocessing

Architecture

Loss

Method A

Use random horizontal flipping for augmentation

Method B

No augmentation

“A **fair comparison** provides evidence that *Method A* is better than *Method B* for some task.”

Preprocessing

Architecture

Loss

Method A

Use a ResNet feature extractor

Method B

Use a VGG-16 feature extractor

“A **fair comparison** provides evidence that *Method A* is better than *Method B* for some task.”



Method A

Use focal loss

Method B

Use cross-entropy loss

What is fair?

Fair: Only one reasonable explanation why Method A $>$ Method B.

Unfair: Several reasonable ways to explain why Method A $>$ Method B.

What if we are comparing fundamentally different methods? E.g. CNN vs Random Forest. There will always be several differences...

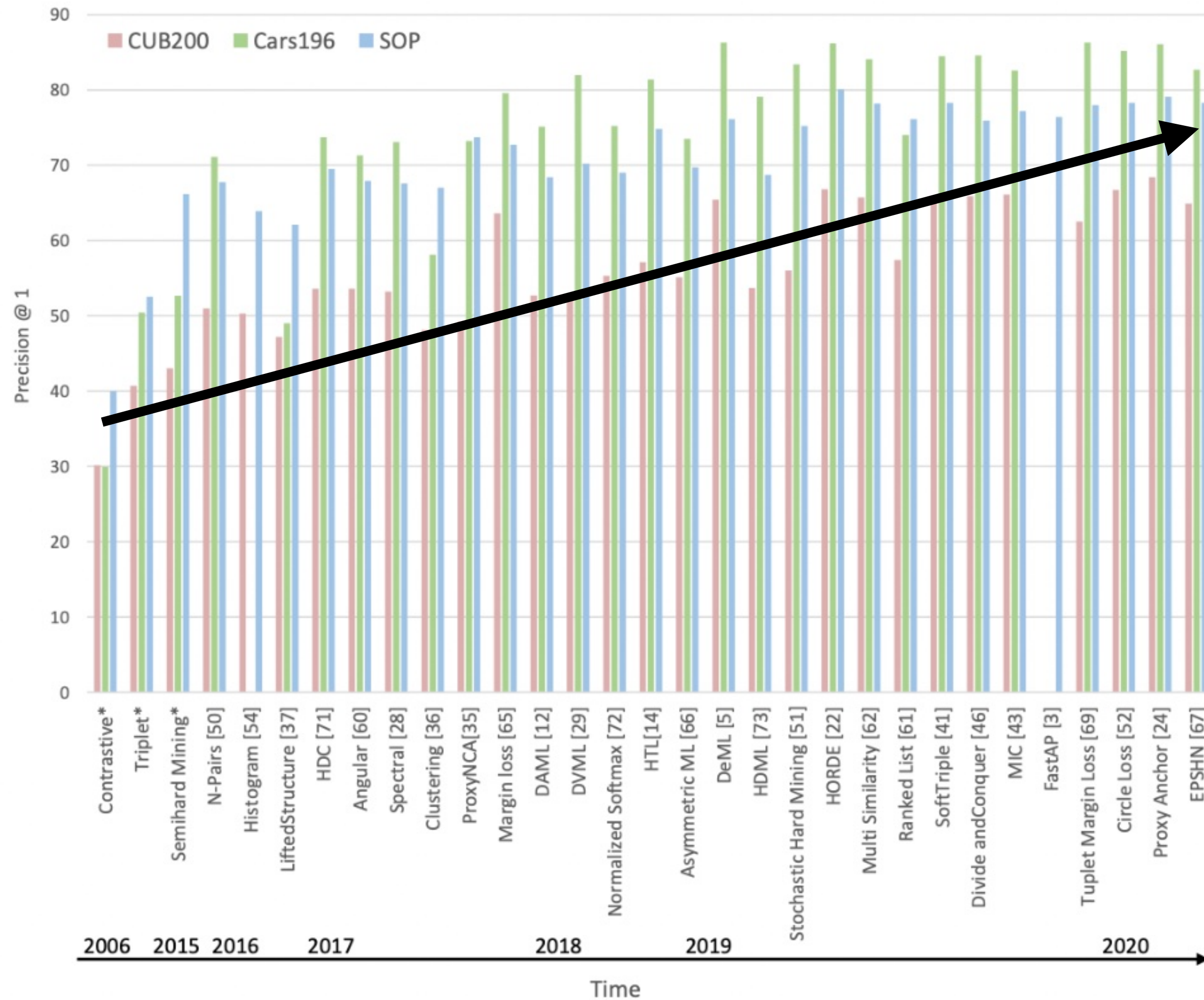
Fair: Method A $>$ Method B, and there is no obvious way to improve the performance of Method B.

Unfair: There are several reasonable ways to improve Method B.

Example: comparing loss functions

Suppose we want to compare two loss functions. What do we keep the same?

- Preprocessing?
- Architecture?
- ~~• Hyperparameter values? (ie batch size, learning rate, etc)~~
- Hyperparameter tuning effort?

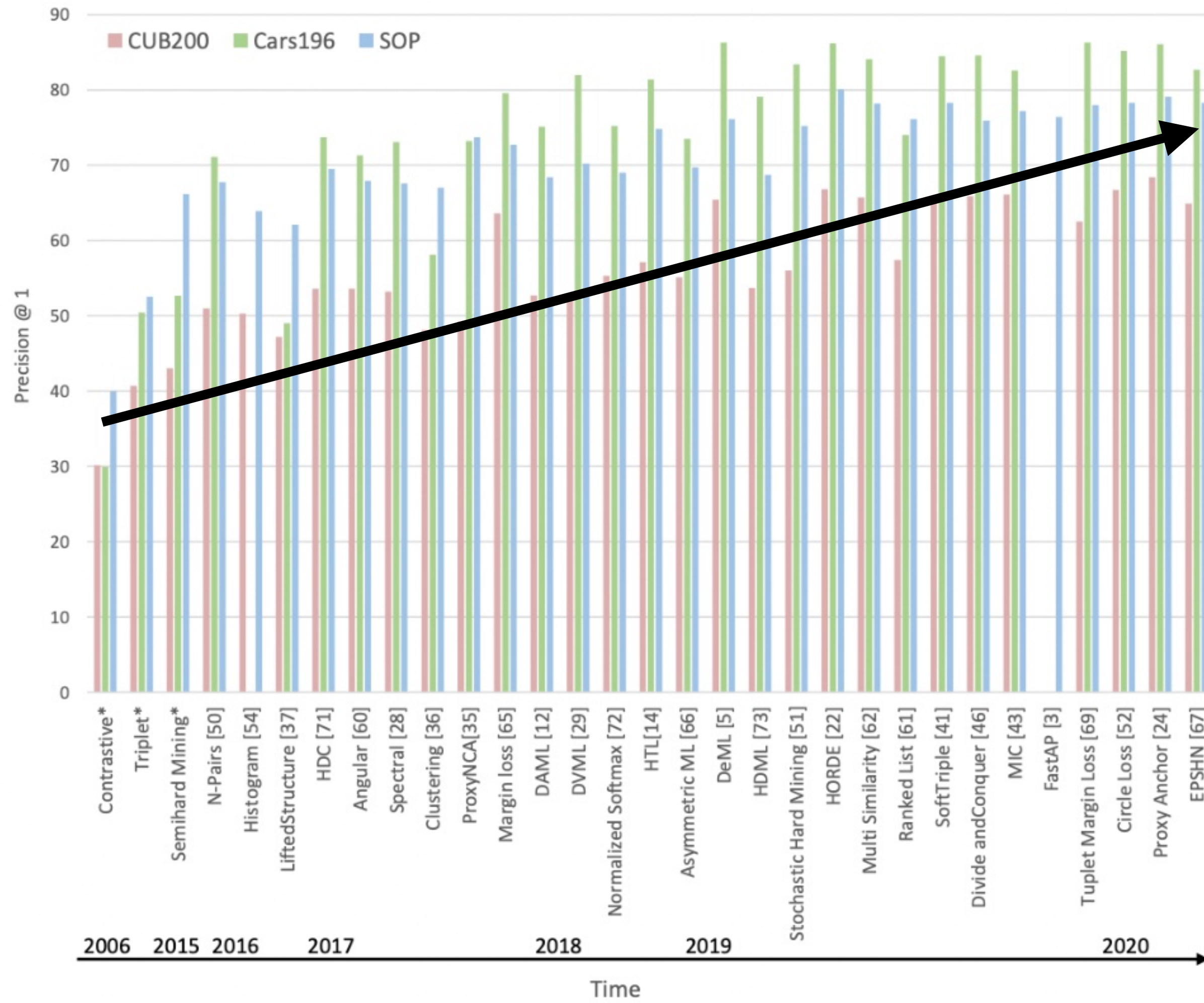


Each group of bars is a different loss function.

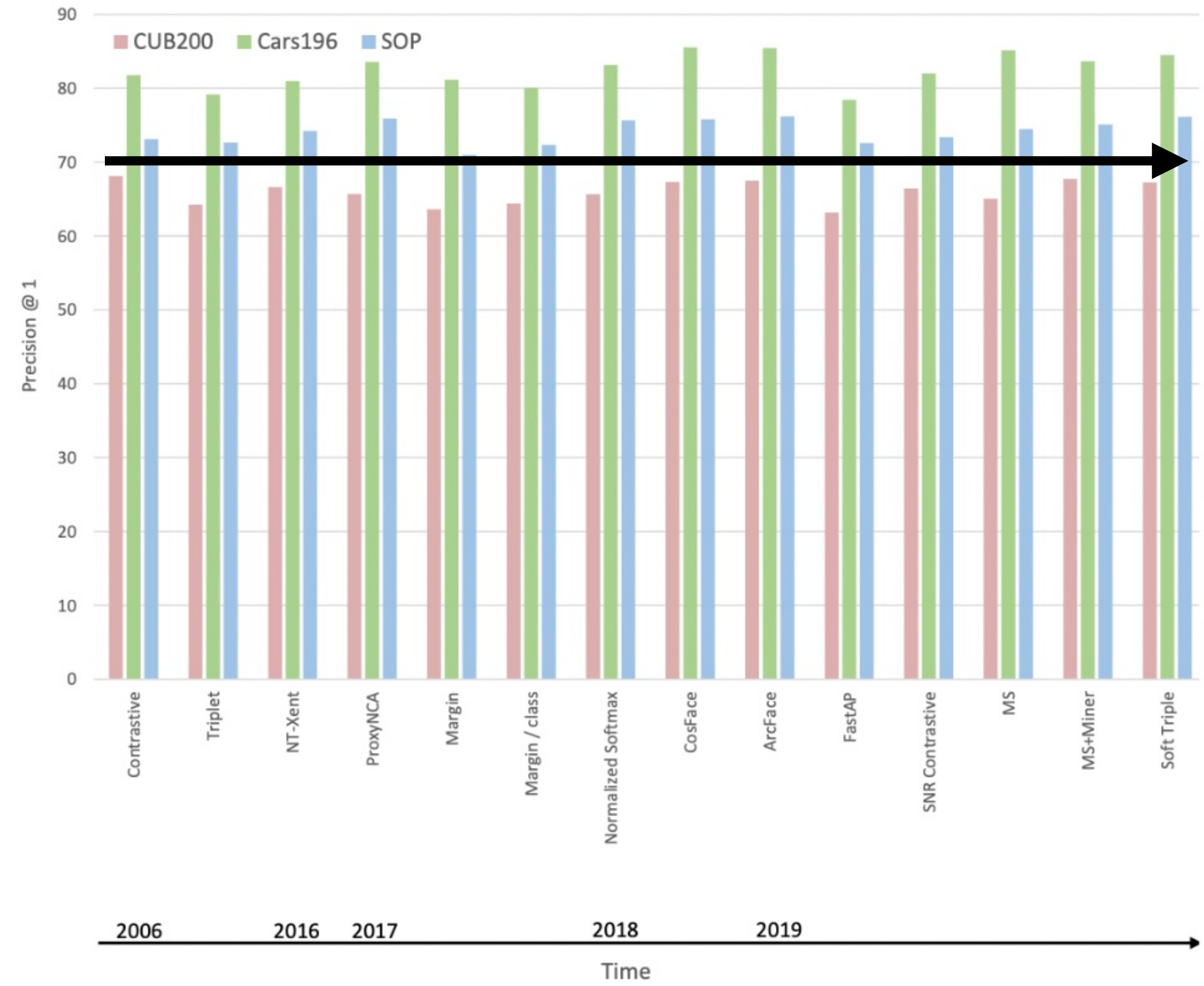
Each color is a different dataset.

(a) The trend according to papers

“One widely-cited paper from 2017 used ResNet-50 and then claimed huge performance gains. This is questionable, because the competing methods used GoogleNet, which has significantly lower initial accuracies. Therefore, much of the performance gain likely came from the choice of network architecture, and not their proposed method.”



(a) The trend according to papers



(b) The trend according to reality

Fair comparison checklist

Suppose we find Method A $>$ Method B

- Are there currently differences that could be controlled for (architecture, training data, etc.)
- Is there a way to improve Method B?
 - Stronger claim: Method A $>$ Method B even if Method B gets an advantage
- Have I invested as much effort tuning Method B as Method A?
- Have I explicitly pointed out any unfairness I couldn't remove?

Ablations

How do we demonstrate progress?

A **fair comparison** provides evidence that *Method A* is better than *Method B* for some task.

An **ablation study** is a set of fair comparisons which show the benefit of each component of a method

Proposed method:

“Start with [baseline], then add [X] and [Y]”

But do we really need both X and Y?

Ablation studies help break down the contribution to performance from each proposed component

Example: Augmentation ablation



(a) Original



(b) Crop and resize



(c) Crop, resize (and flip)



(d) Color distort. (drop)



(e) Color distort. (jitter)



(f) Rotate $\{90^\circ, 180^\circ, 270^\circ\}$



(g) Cutout



(h) Gaussian noise

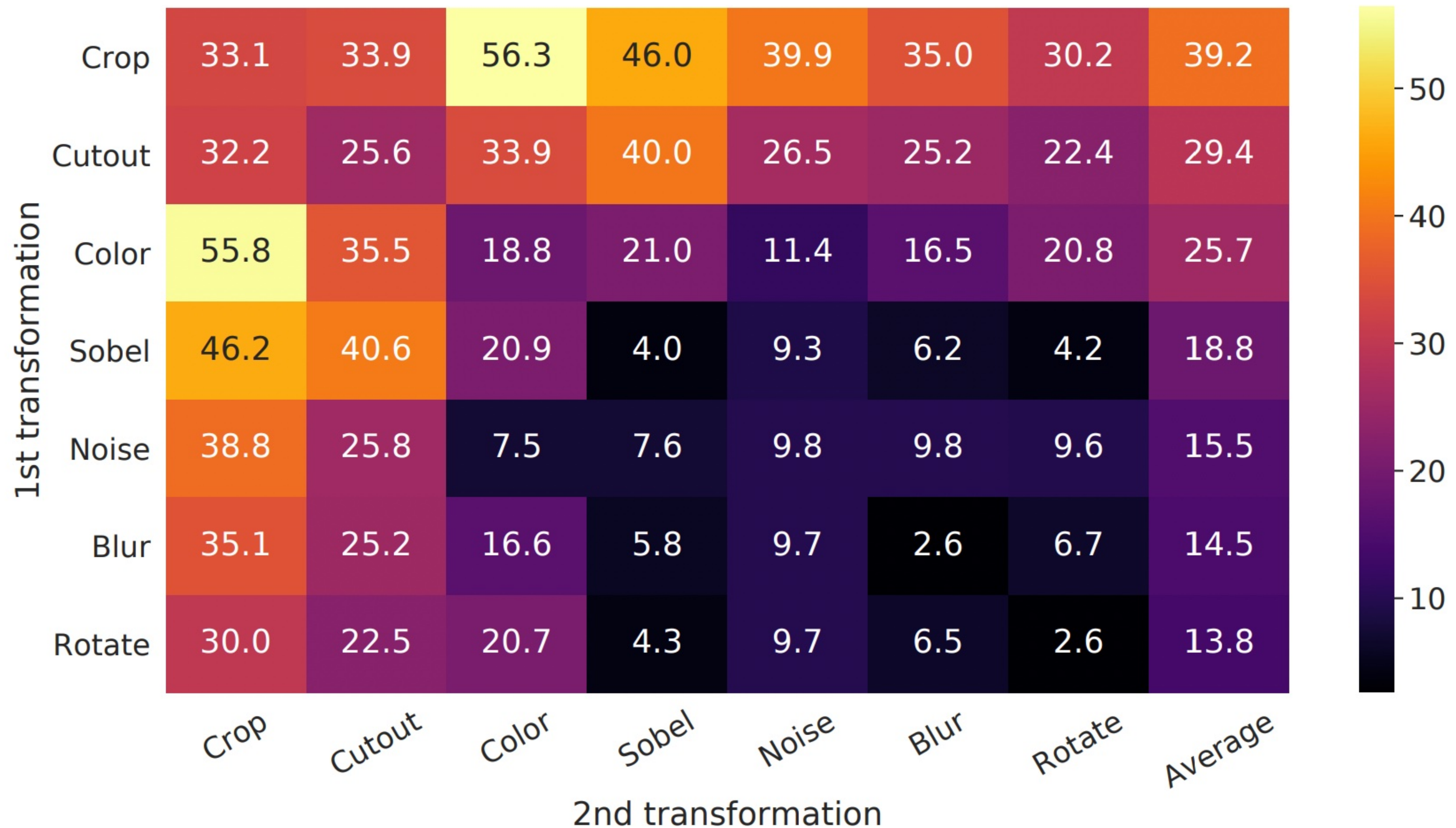


(i) Gaussian blur

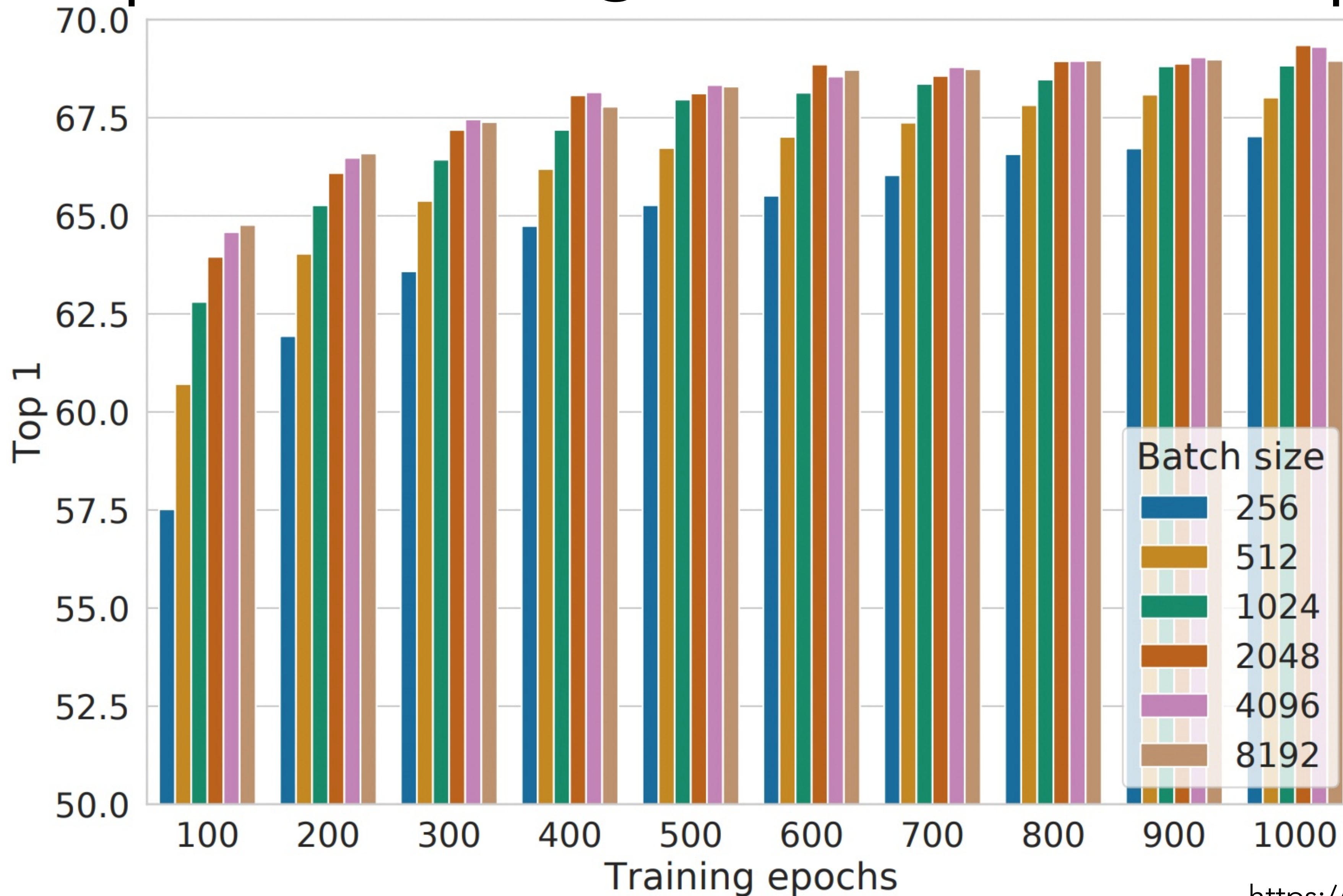


(j) Sobel filtering

Example: Augmentation ablations



Example: Ablating batch size and epochs



Evaluating generative models

Generation is open-ended, which makes evaluation hard


- Test using tests
- Test downstream performance as a proxy
 - Can collect specialized evaluation datasets to enable this
- Test human preferences
- *Learn* human preferences


Testing with tests


Benchmarks like MMLU use multiple choice tests from different disciplines and different levels of difficulty as a mechanism to evaluate generative language models


Professional Law

As Seller, an encyclopedia salesman, approached the grounds on which Hermit's house was situated, he saw a sign that said, "No salesmen. Trespassers will be prosecuted. Proceed at your own risk." Although Seller had not been invited to enter, he ignored the sign and drove up the driveway toward the house. As he rounded a curve, a powerful explosive charge buried in the driveway exploded, and Seller was injured. Can Seller recover damages from Hermit for his injuries?

(A) Yes, unless Hermit, when he planted the charge, intended only to deter, not harm, intruders. 


(B) Yes, if Hermit was responsible for the explosive charge under the driveway. 


(C) No, because Seller ignored the sign, which warned him against proceeding further. 


(D) No, if Hermit reasonably feared that intruders would come and harm him or his family. 


College Mathematics

In the complex z -plane, the set of points satisfying the equation $z^2 = |z|^2$ is a

(A) pair of points 

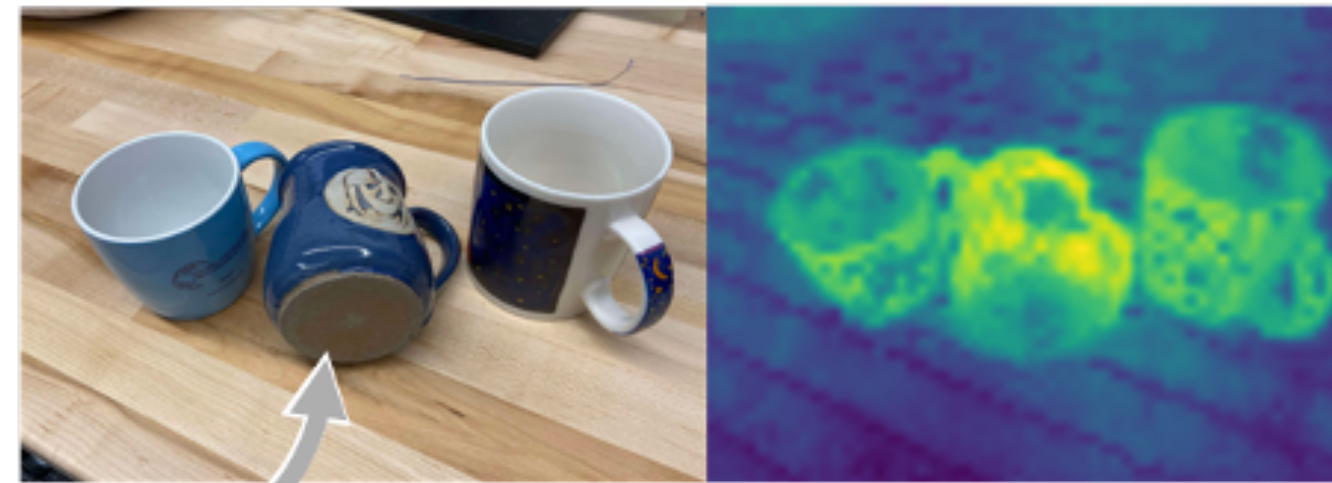
(B) circle 

(C) half-line 

(D) line 

Testing on downstream tasks

Pretrained Representation Space



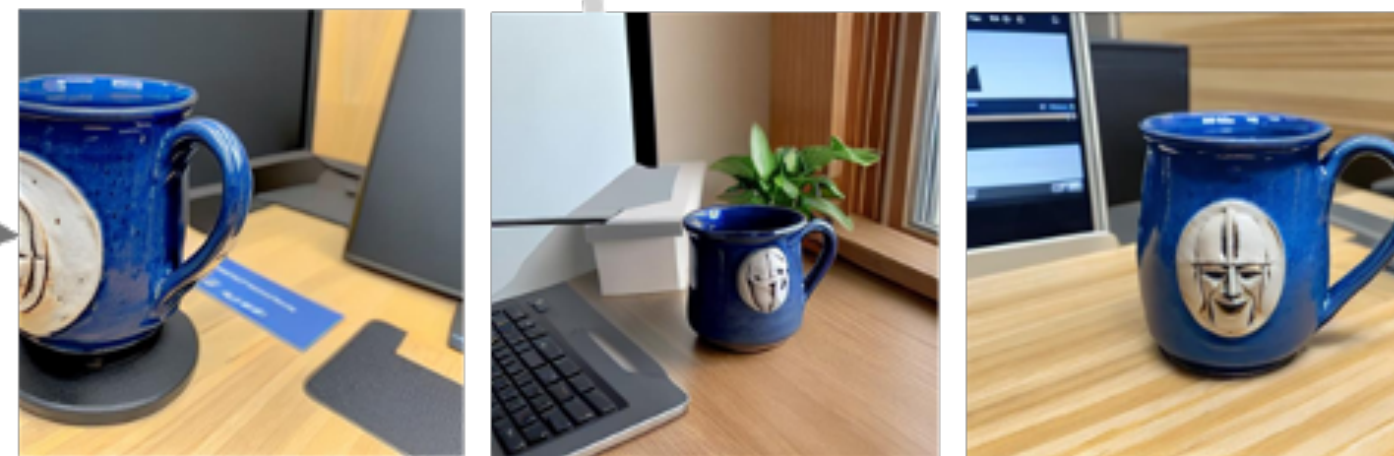
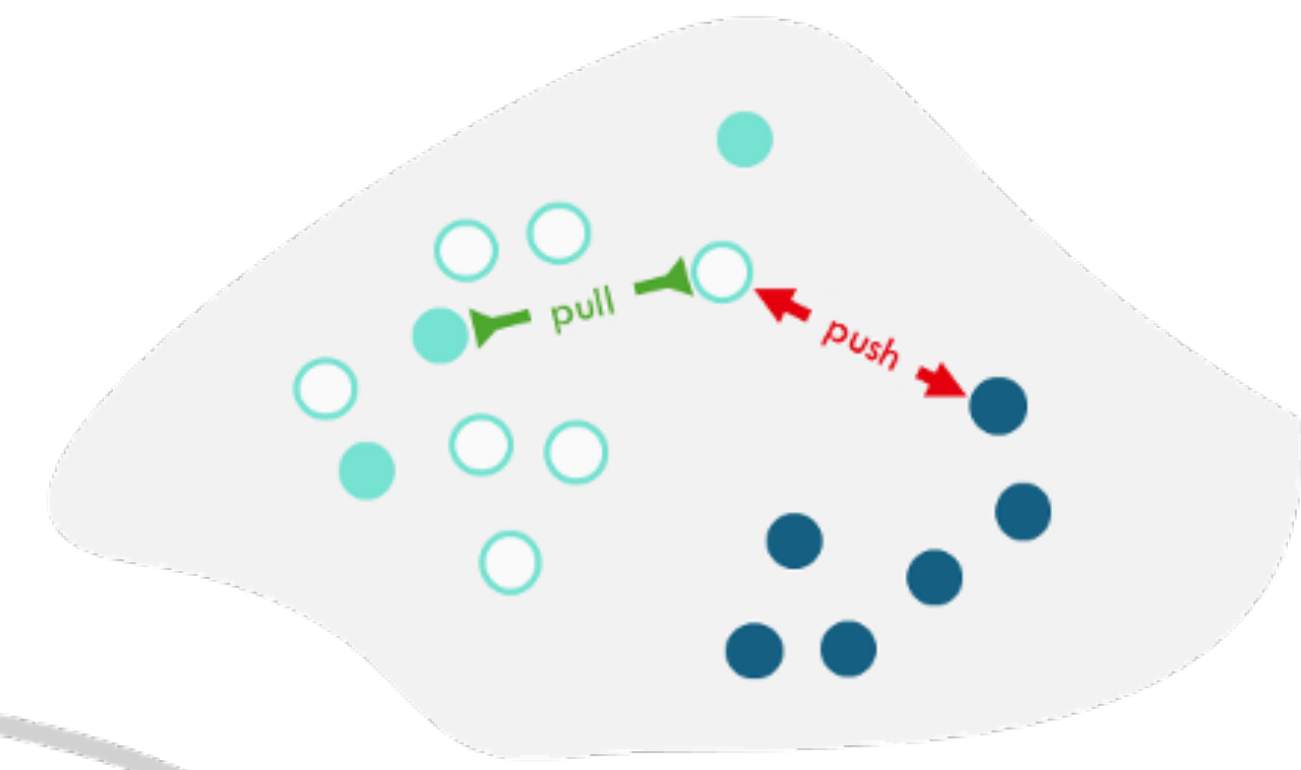
Personalized Representation Space



● Personal Instance Data



Train





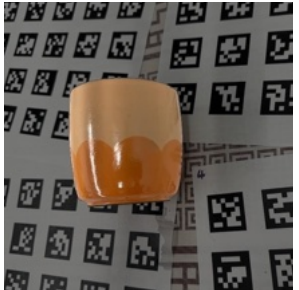
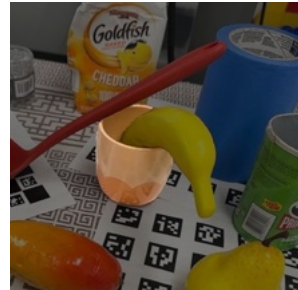
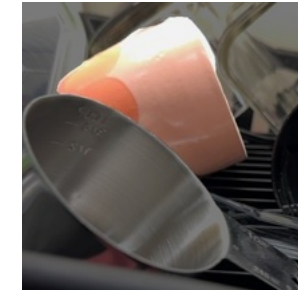









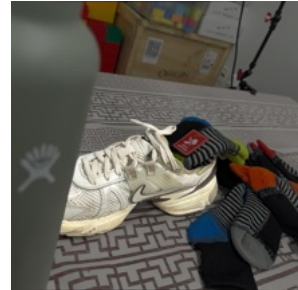










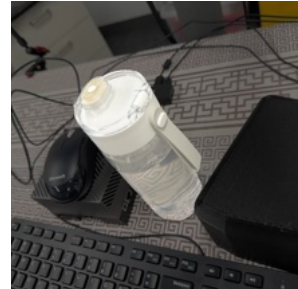









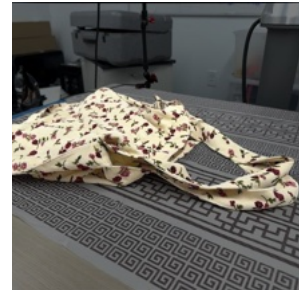








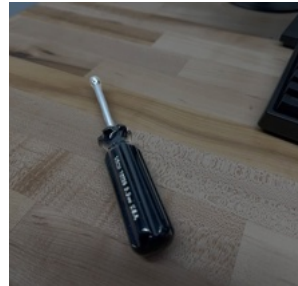



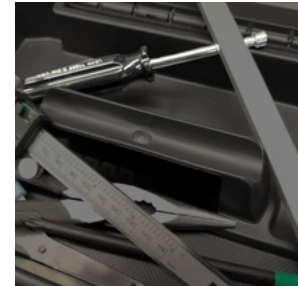



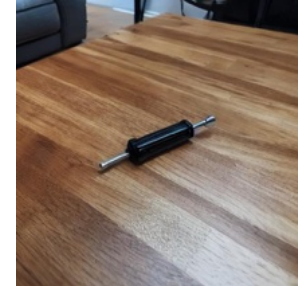

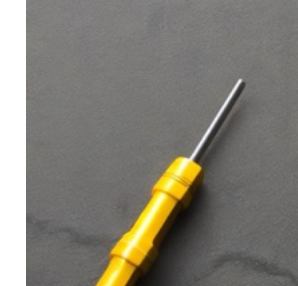
○ Personalized Generated Data

● Other Data (Generated)

Personalized representation from personalized generation

<https://personalized-rep.github.io/>

Collecting targeted evaluation datasets

	PODS Dataset					Generated Images					
	Train	Test ID	Test OOD			DreamBooth Personalized				Negatives	
			Pose	Distractors	Both						
Mug											
Shoe											
Bottle											
Tote											
Screw-driver											

Personalized representation from personalized generation

<https://personalized-rep.github.io/>

Testing via human preference

Many generative models evaluate and compare to prior work via user studies, where humans are asked their preferences between different methods.

User studies are common in social science. Doing them in an unbiased manner takes a lot of care!

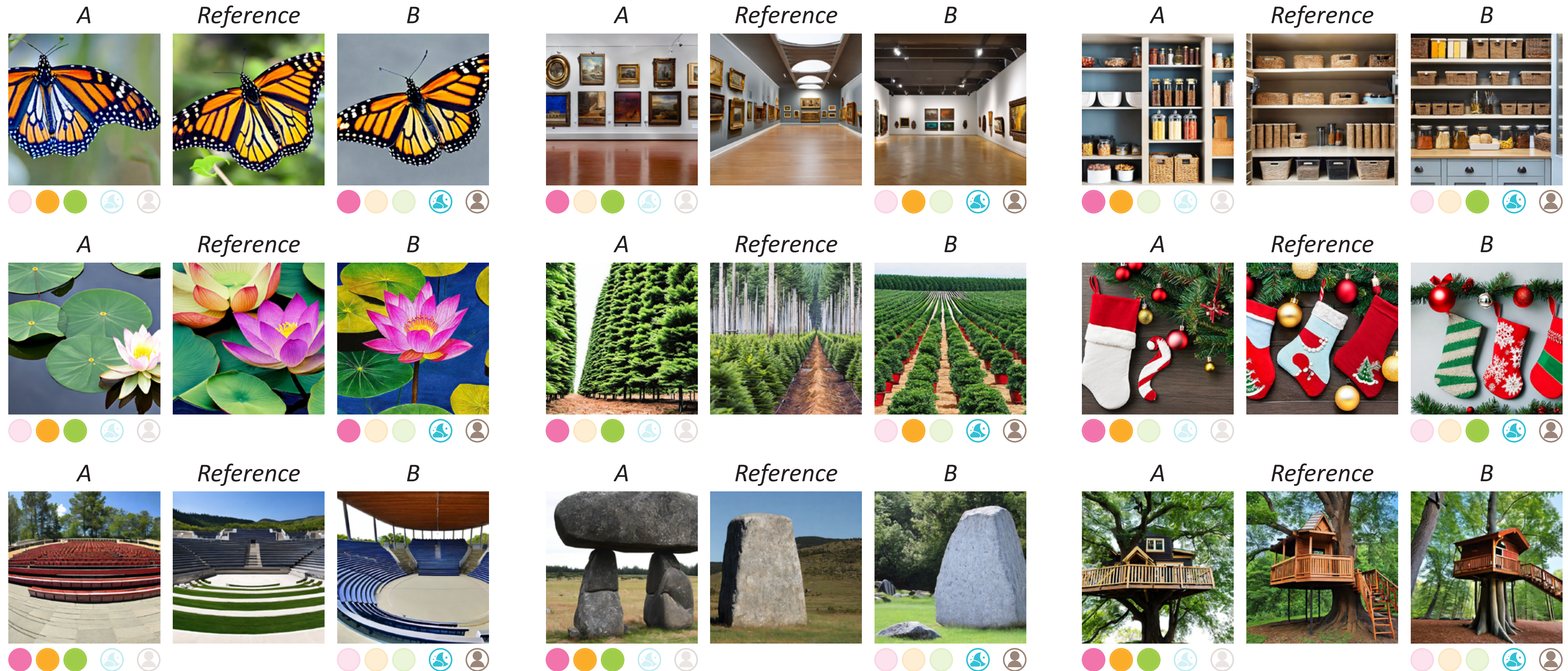
Pick-a-pic is a dataset of pairwise human preferences

<https://arxiv.org/abs/2305.01569>



Figure 1: Images generated via our web application, showing darkened non-preferred images (left) and preferred images (right).

Learned human preference and/or perceptual similarity can also be used as a metric



LPIPS DINO CLIP DreamSim Humans

Testing via human preference







Human preference is a useful signal when evaluating language models as well. But user studies are difficult to scale!

Crowdsourced human preference powers evaluations like Chatbot Arena:

Leaderboard Overview

<https://lmarena.ai/leaderboard>

See how leading models stack up across text, image, vision, and beyond. This page gives you a snapshot of each Arena, you can explore deeper insights in their dedicated tabs. Learn more about it [here](#).

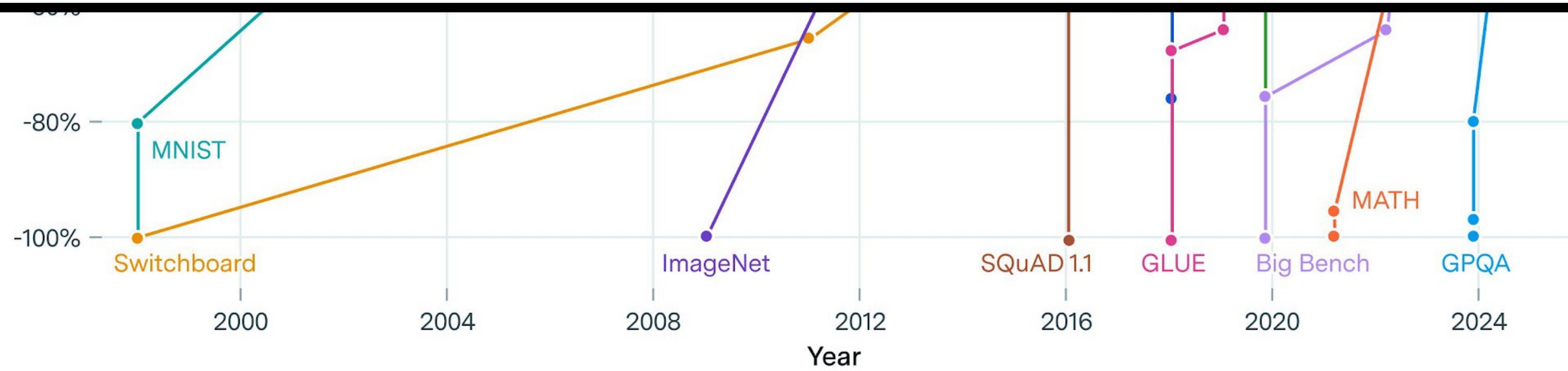
📄 Text 🕒 1 day ago				📁 WebDev 🕒 3 days ago			
Rank ↑↓	Model ↑↓	Score ↓	Votes ↑↓	Rank ↑↓	Model ↑↓	Score ↓ ⓘ	Votes ↑↓
1	 gemini-3-pro	1498 ⓘ	3,768	1	 gemini-3-pro	1487 ⓘ	1,062
2	 grok-4.1-thinking	1483 ⓘ	3,467	2	 gpt-5-medium	1395	2,655
3	 grok-4.1	1464 ⓘ	3,588	3	 claude-opus-4-1-20250805	1394	2,859

"Saturated" Benchmarks

When are benchmarks useful?



Does a saturated benchmark = a solved problem?





Source: International AI Safety Report, Figure 1.4.

Case study: GSM8K (grade school math problems)

ZeroEval: Benchmarking LLMs for Reasoning

Model	Mode	Acc	No answer	Total	Reason Lens
claude-3-5-sonnet-20241022	greedy	96.66	0.00	1319	352.89
gpt-4o-2024-08-06	greedy	96.21	0.00	1319	462.06
o1-mini-2024-09-12	greedy	96.06	0.00	1319	335.77
Llama-3.1-405B-Inst@hyperbolic	greedy	95.98	0.08	1319	421.83
Llama-3.1-405B-Inst@sambanova	greedy	95.91	0.08	1319	464.76
Llama-3.1-405B-Inst-fp8@together	greedy	95.91	0.08	1319	365.07
claude-3-5-sonnet-20240620	greedy	95.60	0.00	1319	465.19
claude-3-opus-20240229	greedy	95.60	0.00	1319	410.62
Mistral-Large-2	greedy	95.53	0.00	1319	391.07

← **Post** Reply


 **Matt Shumer** 
@mattshumer_ Subscribe

I'm excited to announce Reflection 70B, the world's top open-source model.

Trained using Reflection-Tuning, a technique developed to enable LLMs to fix their own mistakes.

405B coming next week - we expect it to be the best model in the world.

Built w/ [@GlaiveAI](#).

Read on :

Benchmark	Reflection 70B	Claude 3.5 Sonnet	Claude 3 Opus	GPT-4o	Gemini 1.5 Pro	Llama 3.1 405B
GPQA	55.3% (0-shot Reflection)	59.4%* (0-shot CoT)	50.4% (0-shot CoT)	53.6% (0-shot CoT)	—	50.7% (0-shot)
MMLU	89.9% (0-shot Reflection)	88.7%** (5-shot) 88.3% (0-shot CoT)	86.8% (5-shot) 85.7% (0-shot CoT)	88.7% (0-shot CoT)	85.9% (5-shot)	87.3% (5-shot) 88.6% (0-shot CoT)
HumanEval	91% (0-shot Reflection)	92.0% (0-shot)	84.9% (0-shot)	90.2% (0-shot)	84.1% (0-shot)	89.0% (0-shot)
MATH	79.7% (0-shot Reflection)	71.1% (0-shot CoT)	60.1% (0-shot CoT)	76.6% (0-shot CoT)	67.7% (4-shot)	73.8% (0-shot CoT)
GSM8K	99.2% (0-shot Reflection)	96.4% (0-shot CoT)	95.0% (0-shot CoT)	—	90.8% (11-shot)	96.8% (8-shot CoT)
IFEval	90.13% (0-shot Reflection)	88.0%	—	85.6%	—	88.6%



Hugh Zhang

@hughbzhang



Hey Matt! This is super interesting, but I'm quite surprised to see a GSM8k score of over 99%. My understanding is that it's likely that more than 1% of GSM8k is mislabeled (the correct answer is actually wrong)!

3:34 PM · Sep 5, 2024 · **28.5K** Views



Post



Jing Yu Koh

@kohjingyu



On GSM8K, 98% is better than 99%.

8:54 AM · Sep 6, 2024 · **13.9K** Views



Peter Henderson ✓

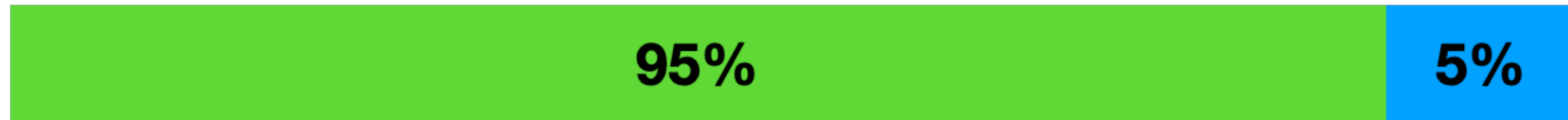
@PeterHndrsn



Has anyone analyzed what's the ground truth label error rate for GSM8k? It's possible we entered data leakage and overfitting territory a while ago. x.com/polynoamial/st...

The community's intuition is that 100% isn't good, because some questions are bad or wrong

→ World 1: saturated = reliability



All label noise



→ World 2: saturated \neq reliability



Some label noise



But also real, lingering model failures!



Let's figure it out!

We cleaned up subsets of 15 popular benchmarks with a custom labeling tool

DATASET: gsm8k QUESTION # 183 of 300 VALIDATED: 77/300 [← Previous](#) [Next →](#)

QUESTION Don't wrap question

Solve the following math word problem.

Jean is two years older than Mark. Two years ago Mark was 5 years older than half Jan's age. If Jan is 30 how old is Jean?

Think step-by-step. Then, provide the final answer as a single integer in the format "Answer: XXX" with no extra formatting.

ANSWER

23

PREDICTIONS

Meta-Llama-3.1-405B-Instruct 23 ✓ == label == user answer	claude-3-5-sonnet 23 ✓ == label == user answer	o1-mini 23 ✓ == label == user answer	gpt-4o-2024-11-20 24 ! != label	gpt-4o-2024-08-06 24 ! != label	gemini-1.5-pro 23 ✓ == label == user answer	Meta-Llama-3.1-70B-Instruct 23 ✓ == label == user answer
gpt-4o-mini 23 ✓ == label == user answer	gemini-1.5-flash 23 ✓ == label == user answer	mistral-small 23 ✓ == label == user answer	o1-preview-2024-09-12 23 ✓ == label == user answer	Llama-3.3-70B-Instruct 23 ✓ == label == user answer	grok-2-1212 23 ✓ == label == user answer	claude-3-5-sonnet-20241022 23 ✓ == label == user answer
gemini-2.0-flash 30 ! != label	deepseek/deepseek-chat 23 ✓ == label == user answer	Qwen/Qwen2.5-72B-Instruct 23 ✓ == label == user answer	o1-2024-12-17-high 24 ! != label	o1-2024-12-17-med 24 ! != label	deepseek-r1 023 ✓ == label	o3-mini-2025-01-31-high 23 ✓ == label == user answer

! At least one model prediction does not match the answer.

PREDICTION EXPLANATION

Let's break down the problem step by step.

- Jan is 30 years old.
- Two years ago, Jan was $30 - 2 = 28$ years old.
- Two years ago, Mark was 5 years older than half Jan's age. Half of Jan's age two years ago is $28 / 2 = 14$. So, Mark was $14 + 5 = 19$ years old two years ago.
- Mark's current age is $19 + 2 = 21$ years old.
- Jean is two years older than Mark, so Jean's current age is $21 + 2 = 23$ years old.

Answer: 23

USER ANSWER

23

✓ Matches existing answer

BAD QUESTION? [SPACE]

[\[f\] Go to next incorrect prediction](#)
[\[g\] Go to next incorrect + unvalidated](#)
[\[h\] Go to next incorrect prediction from the selected model](#)
[\[j\] Go to next unvalidated prediction from the selected model](#)

(a) Mislabeled question, *SVAMP*

Question: You had 14 bags with equal number of cookies. If you had 28 cookies and 86 candies in total, how many bags of cookies do you have?

Solution: 2

There are 14 bags, not 2.

(c) Ambiguity, *VQA v2.0*



Question: Does the baby have socks on?²

There is no way to tell.

(b) Logical contradiction, *GSM8K*

Ten stalls have 20 cows each. Mr. Syllas buys 40 cows and divides them equally, putting an equal number of the new cows into each of the twenty stalls. How many cows are in 8 of the stalls?

There are both ten and twenty stalls.

(d) Clear flaw / ill-posed, *MMLU HS Math*

A curve is given parametrically by the equations

Options:

A) $\pi/2$ B) π C) $2 + \pi$ D) 2π

The equations for the curve are missing.

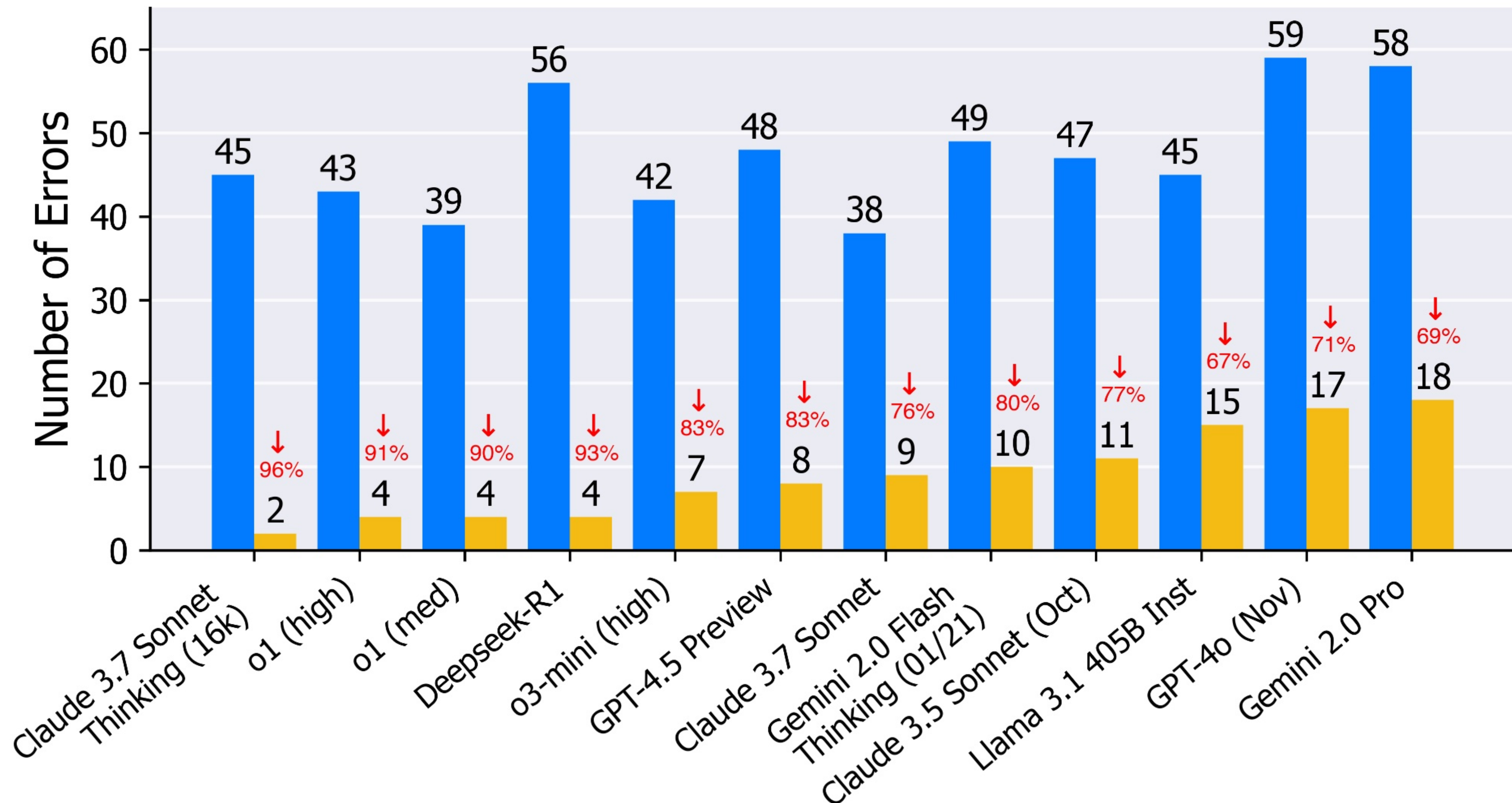
Almost 10% of GSM8K

Up to 34% of question for reading comprehension benchmarks

Over 50% of VQA

	<i>SingleOp</i>	<i>SingleEq</i>	<i>MultiArith</i>	<i>SVAMP</i>	<i>GSM8K</i>	<i>MMLU HS Math</i>	<i>Logic Ded. 3-Obj</i>	<i>Object Counting</i>	<i>Navigate</i>	<i>TabFact</i>	<i>HotpotQA</i>	<i>SQuAD2.0</i>	<i>DROP</i>	<i>Winograd WSC</i>	<i>VQA v2.0</i>
Type	Math					Logic				Tab	RC			CR	Vis
# Original Questions	159	109	174	300	300	270	200	200	200	200	250	250	250	200	600
<i>Platinum Labeling</i>															
# Bad Questions	9	9	3	27	26	2	0	9	0	27	66	86	41	5	352
# Mislabeled	0	0	3	3	1	0	0	0	0	3	6	4	6	0	—
# Platinum Questions	150	100	171	273	274	268	200	191	200	173	184	164	209	195	248

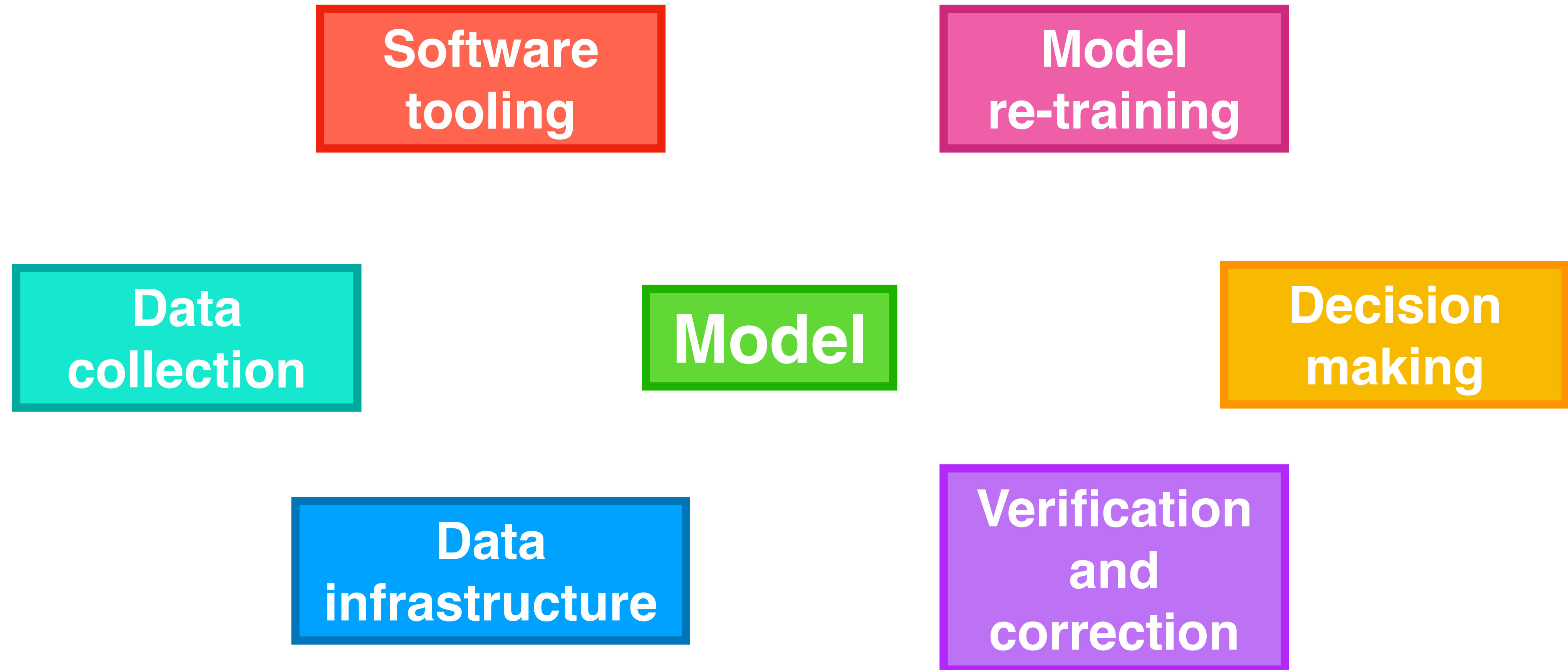
GSM8K Test Errors



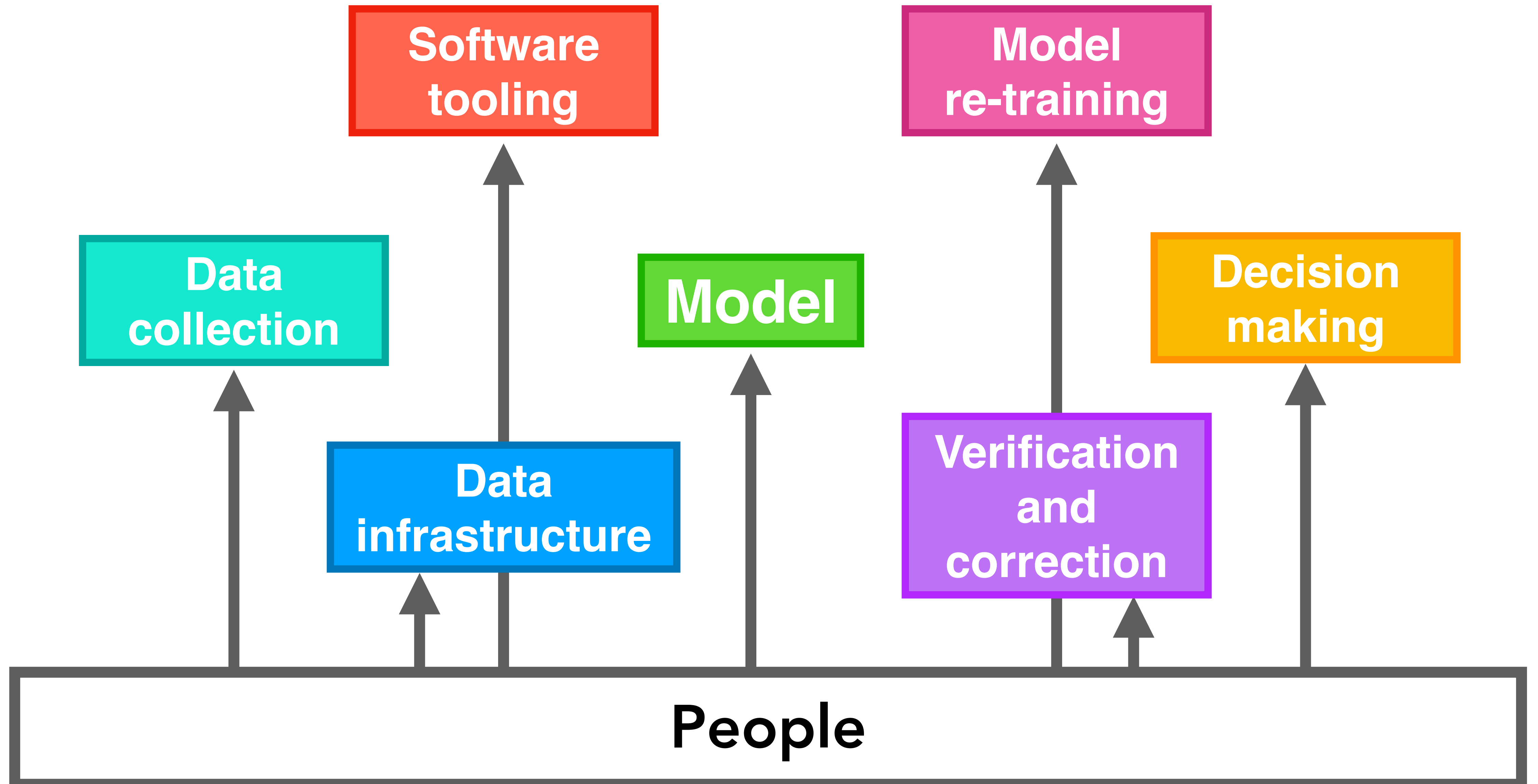
Good performance on realistic benchmarks \neq impact

Model

Good performance on realistic benchmarks \neq impact



Good performance on realistic benchmarks \neq impact



Where do models fail?

Adversarial Examples

Adversarial examples

“pig”



91% confidence

Adversarial examples

“pig”

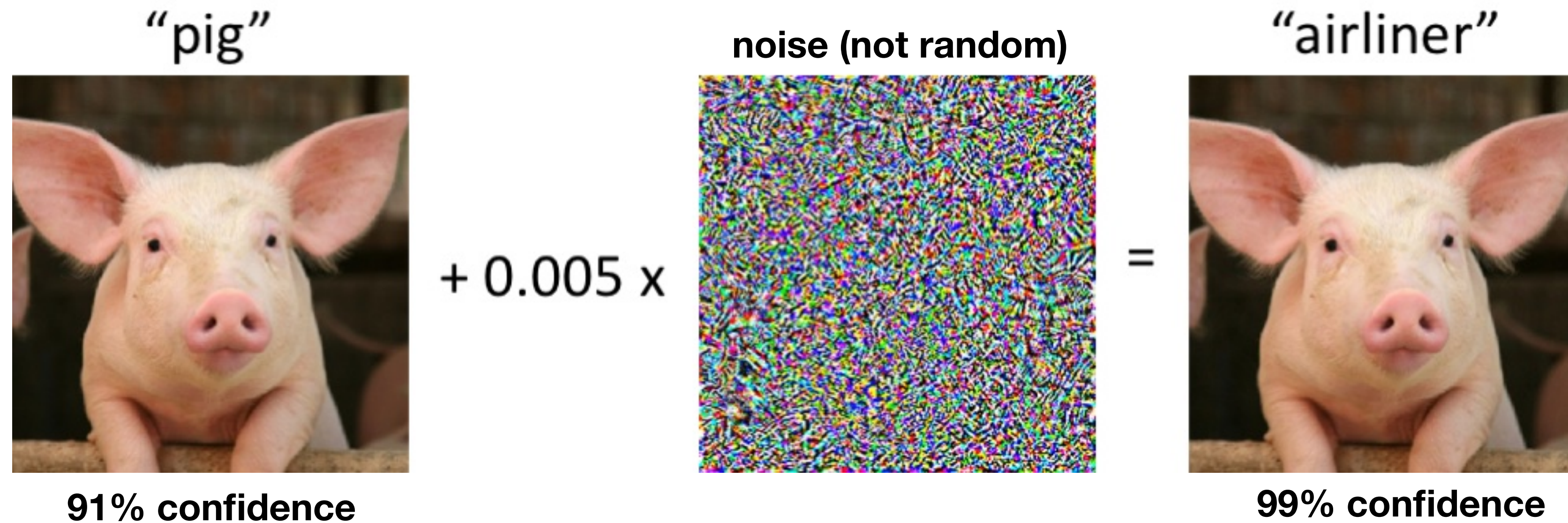


+ 0.005 x



91% confidence

Adversarial examples

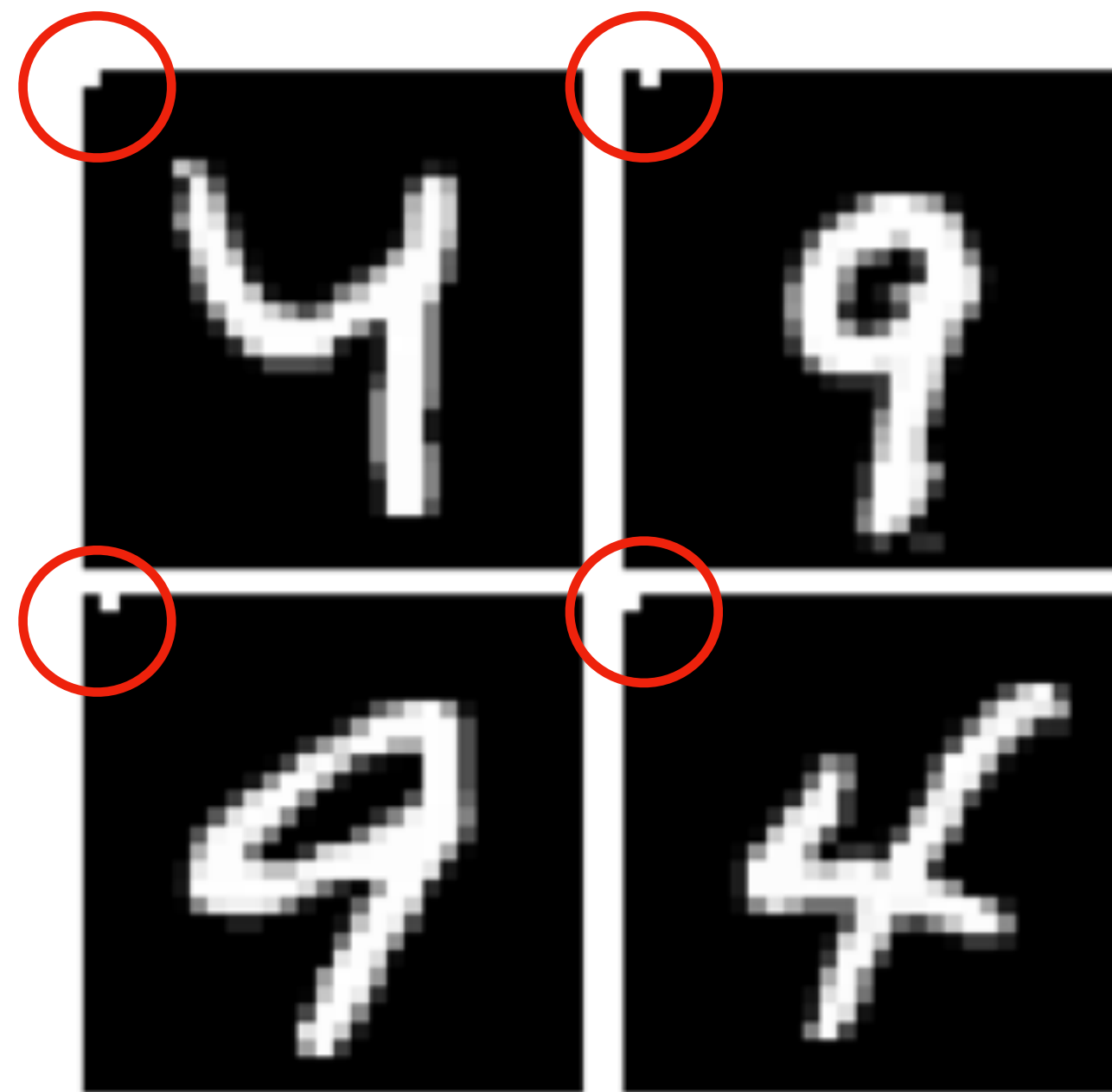


- ML model predictions are (mostly) accurate but can be brittle

What do adversarial examples tell us?

What do adversarial examples tell us?

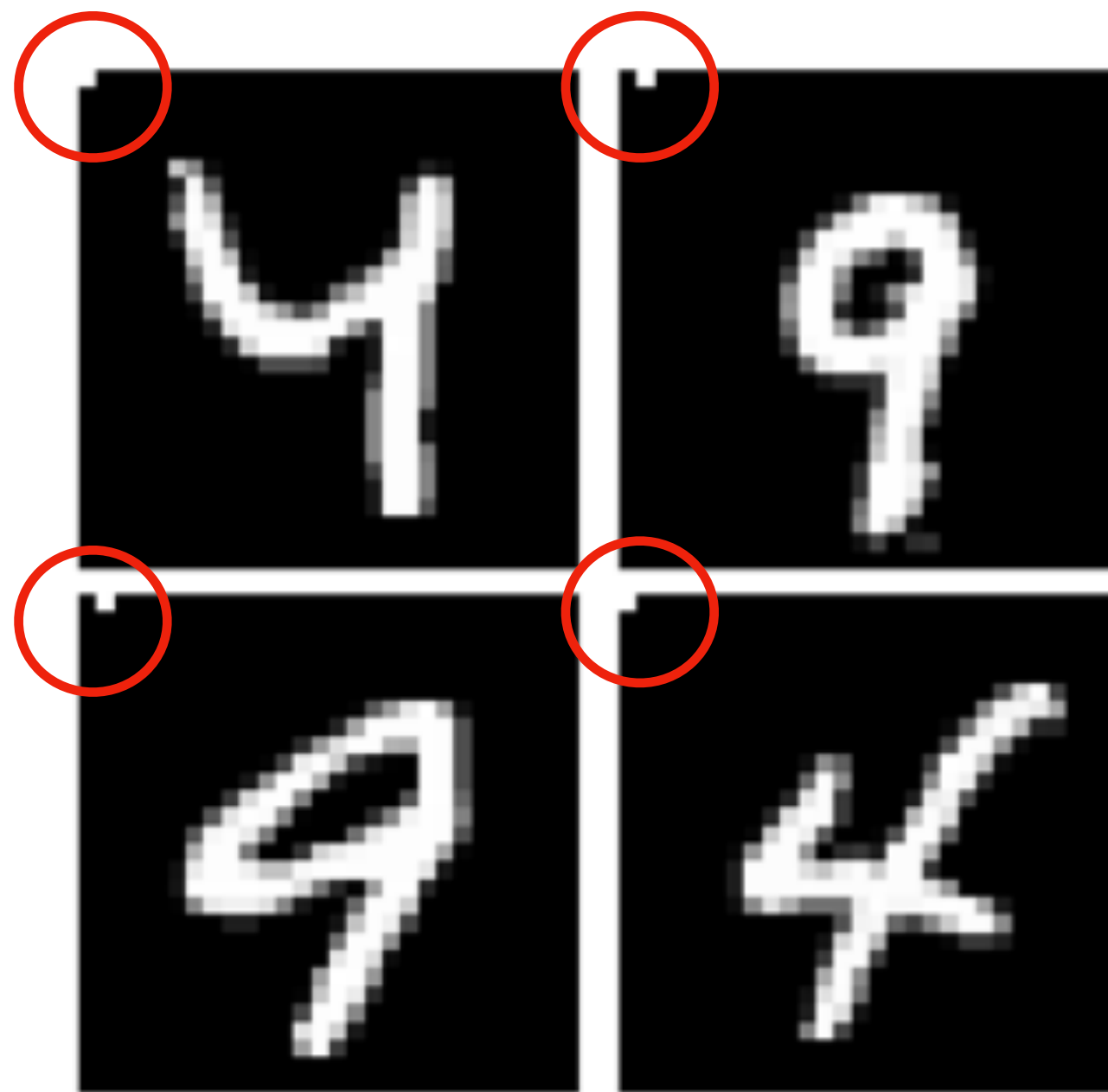
- something about the input “features” that are critical for the model’s decision
- Example:



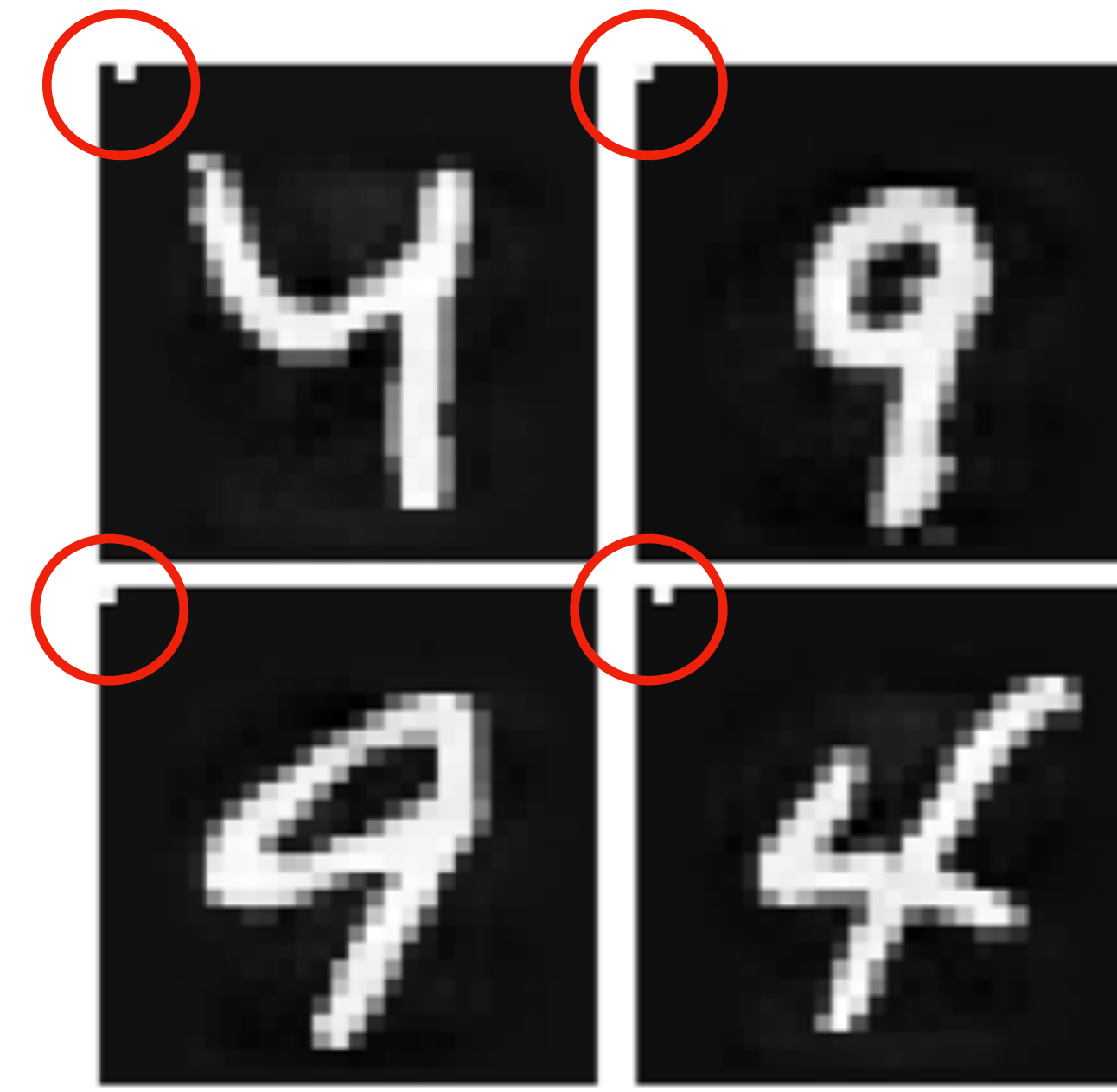
**Training data:
classify 4 vs 9**

What do adversarial examples tell us?

- something about the input “features” that are critical for the model’s decision
- Example:

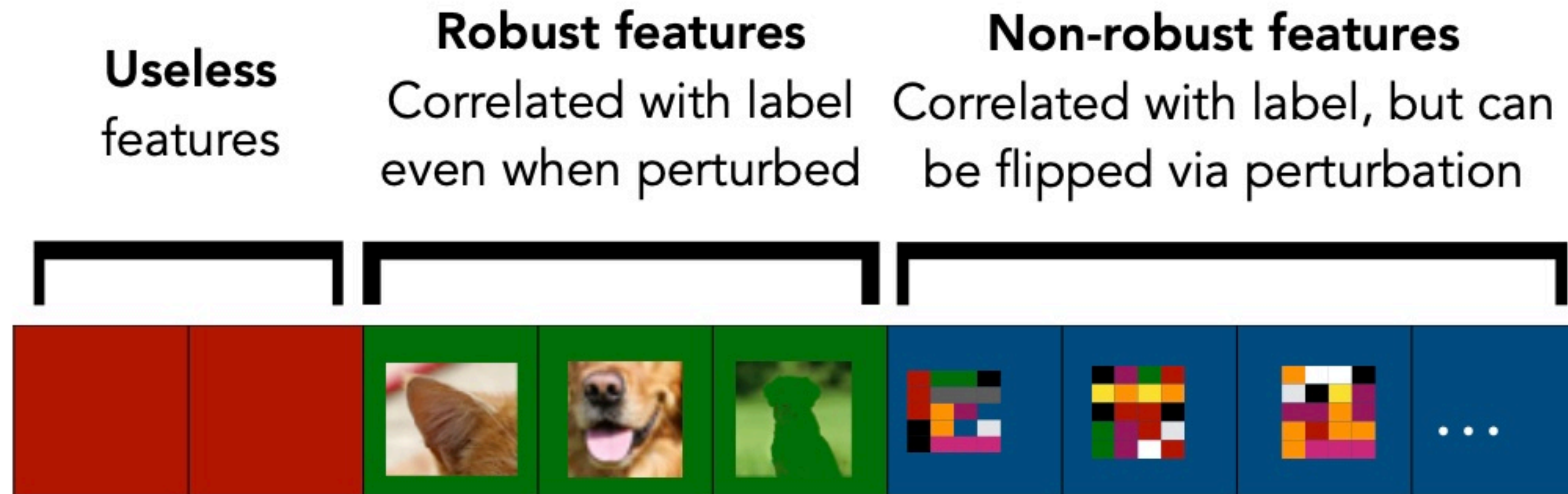


**Training data:
classify 4 vs 9**



**Adversarial
perturbations**

Predictive features



- Many features may be **correlated with the label** and hence predictive and help with accuracy, *beyond what humans would use*.

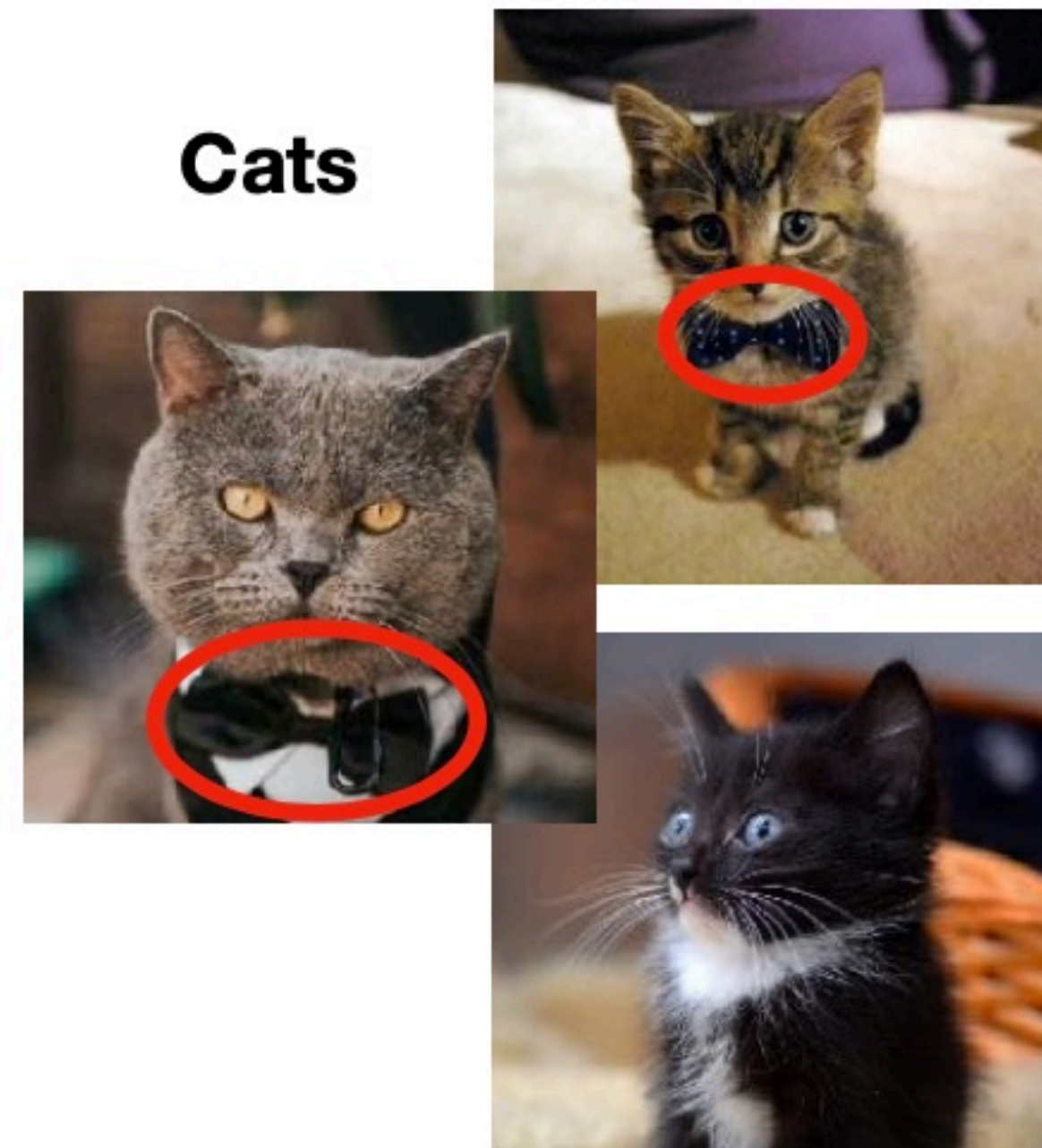
Where do these correlations come from?

- Data

Dogs



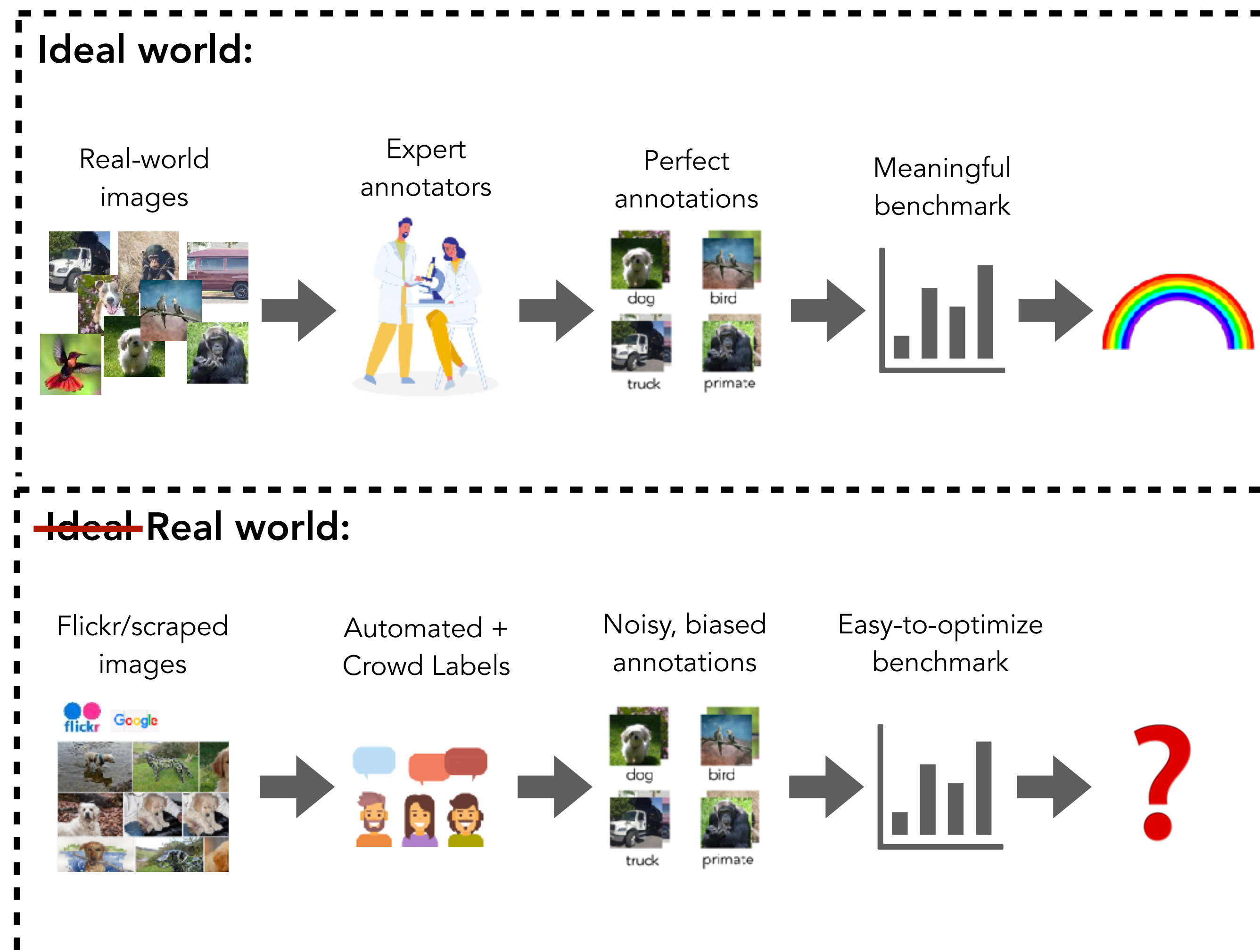
Cats



“Fish” from the ImageNet training set

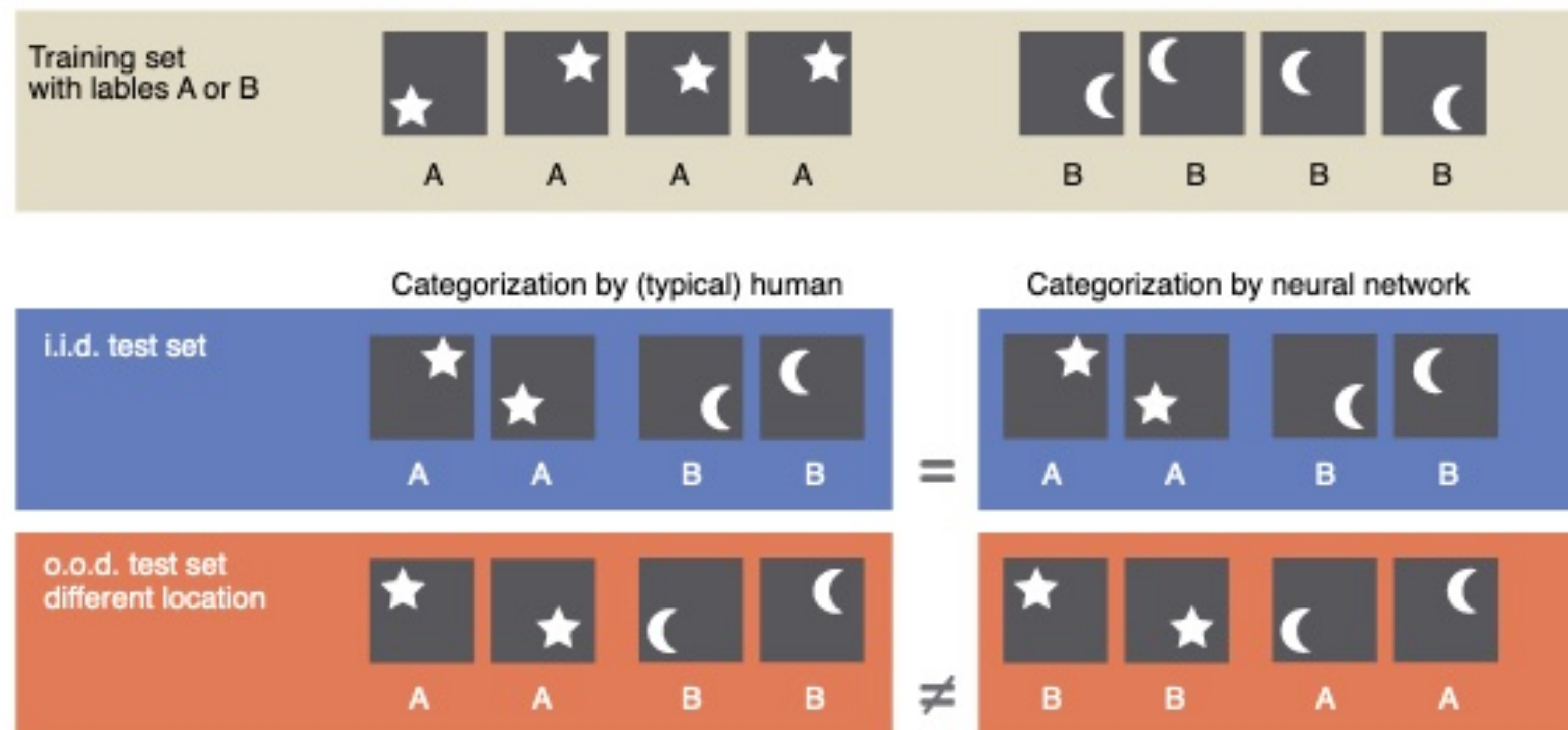
Where do these correlations come from?

- ...and how we create datasets



It's all "shortcuts"

- Shortcuts: features correlated with label in the training data, but not under realistic distribution shifts
- Models will use them and not generalize if features are no longer correlated



It's all "shortcuts"

- Shortcuts: features correlated with label in the training data, but not under realistic distribution shifts
- Models will use them and not generalize if features are no longer correlated
- This is related to **data**, not models: ***adversarial examples transfer across models trained on the same dataset***

What can these shortcuts look like?



A herd of sheep grazing on a lush green hillside
Tags: grazing, sheep, mountain, cattle, horse



NeuralTalk2: A flock of birds flying in the air
Microsoft Azure: A group of giraffe standing next to a tree
Image: Fred Dunn, <https://www.flickr.com/photos/gratapictures> - CC-BY-NC

What can these shortcuts look like?



A herd of sheep grazing on a hillside
Tags: grazing, sheep, mountains

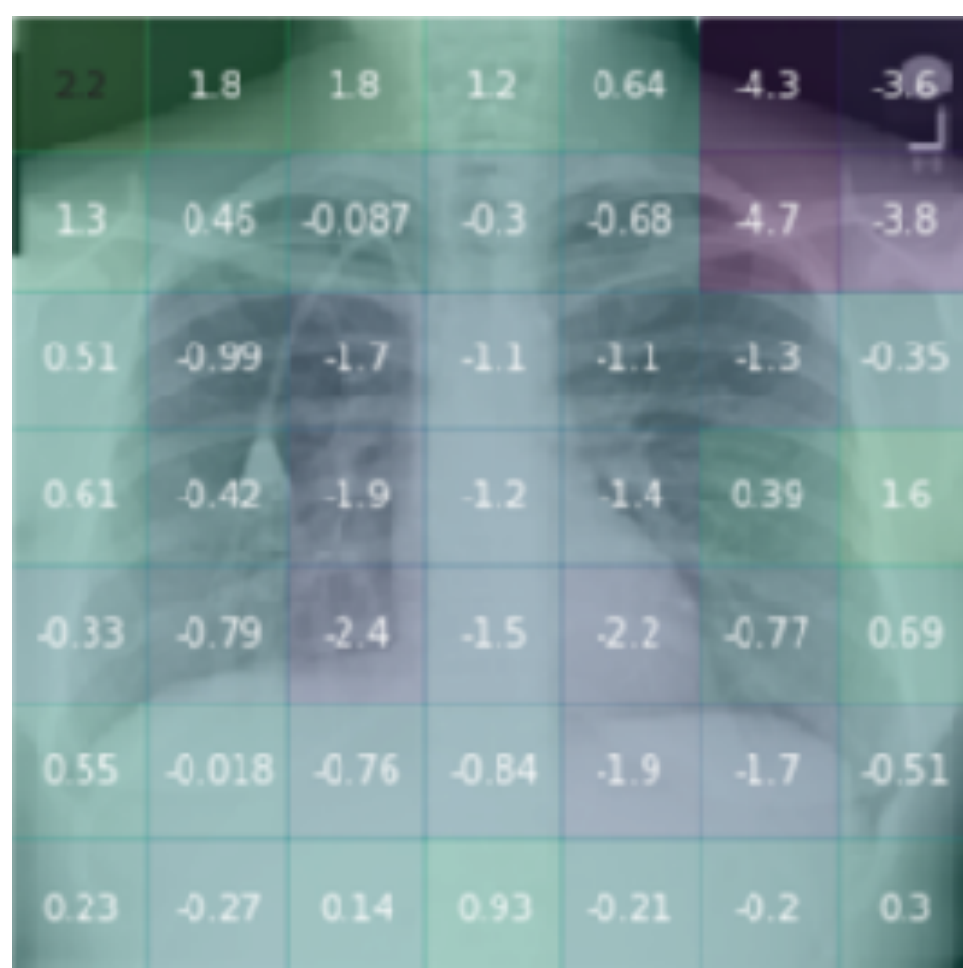


A flock of birds flying in the air
A group of giraffe standing next to a tree
www.flickr.com/photos/gratapictures - CC-BY-NC



Left: A man is holding a dog in his hand
Right: A woman is holding a dog in her hand
Image: @SouperSarah

What can these shortcuts look like?



"...if an image had a ruler in it, the algorithm was more likely to call a tumor malignant..."

[Esteva et al. 2017]

"CNNs were able to detect where an x-ray was acquired [...] and calibrate predictions accordingly."

[Zech et al. 2018]



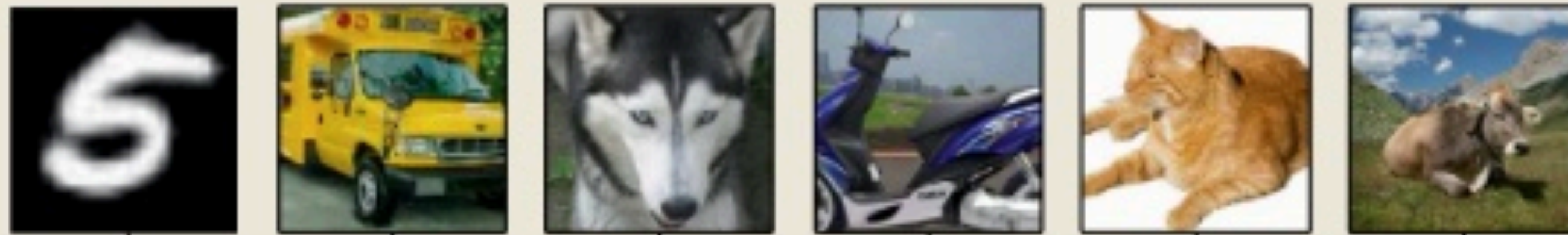
not all predictive patterns are desirable

Many more...

Same category for humans
but not for DNNs (intended generalization)

Same category for DNNs
but not for humans (unintended generalization)

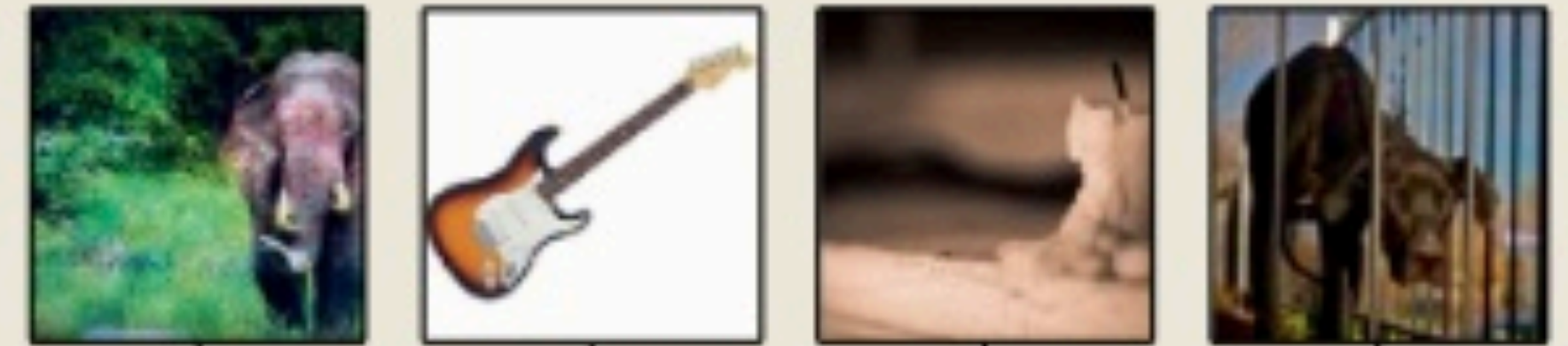
i.i.d.



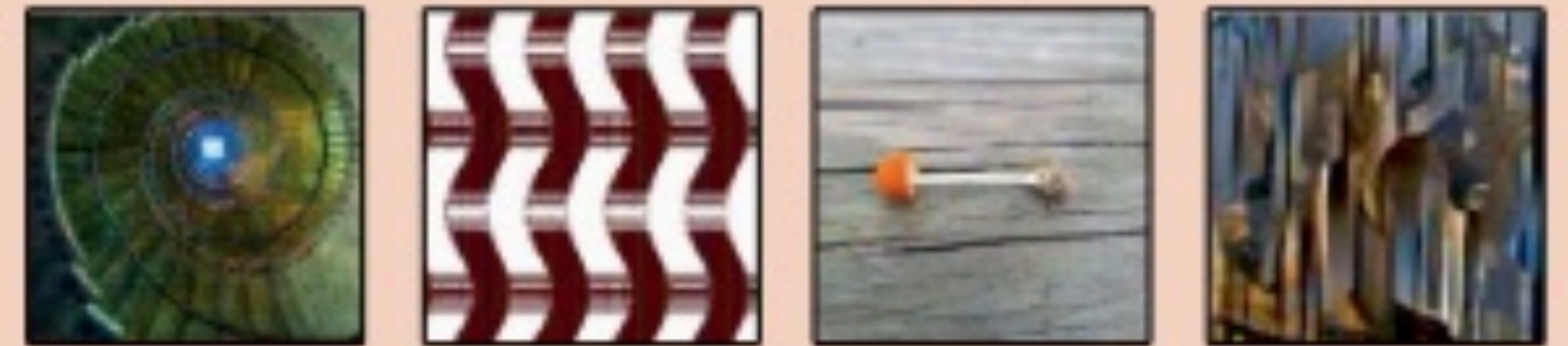
Domain shift Wang 2018 Adversarial examples Szegedy 2013 Distortions Dodge 2019 Pose Alcorn 2019 Texture Geirhos 2019 Background Beery 2018



o.o.d.



Excessive invariance Jacobson 2019 Fooling images Nguyen 2015 Natural adversarials Hendrycks 2019 Texturized images Brendel 2019



Transformers Learn Shortcuts to Automata

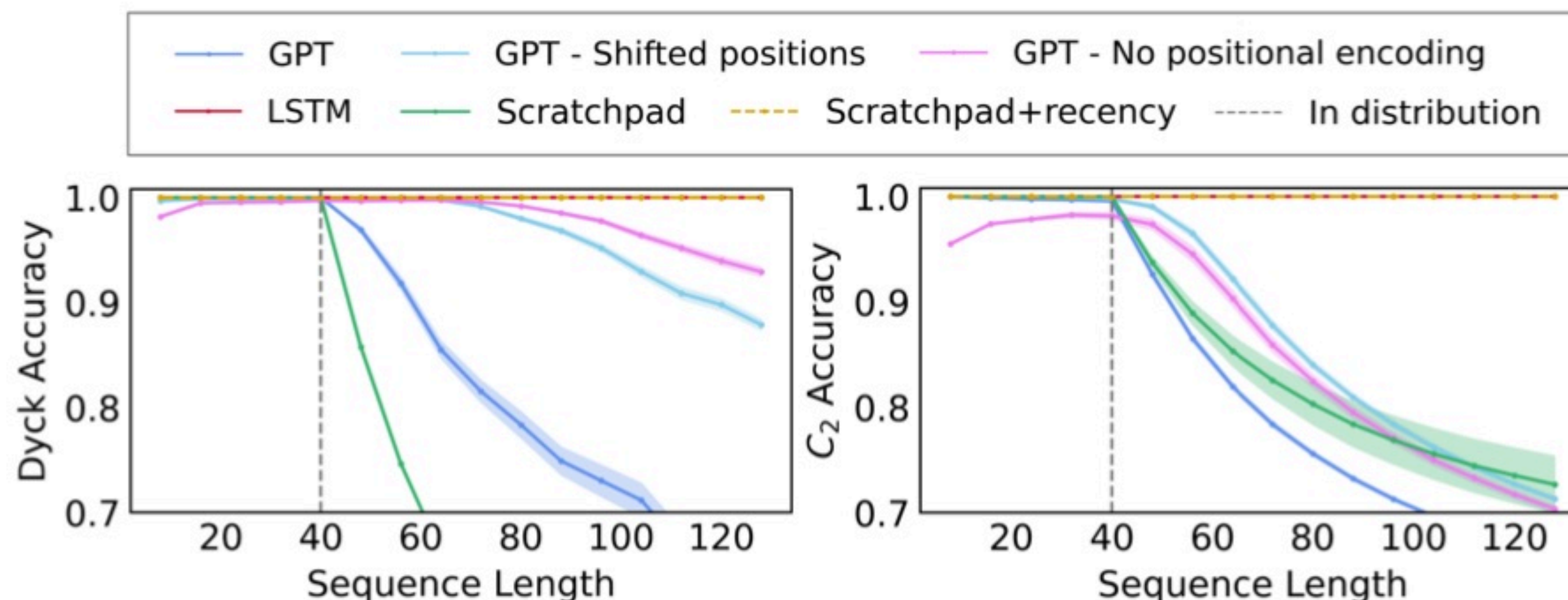
Bingbin Liu^{1*} Jordan T. Ash² Surbhi Goel^{2,3} Akshay Krishnamurthy² Cyril Zhang²

¹Carnegie Mellon University ²Microsoft Research NYC ³University of Pennsylvania
bingbinl@cs.cmu.edu, {ash.jordan, goel.surbhi, akshaykr, cyrilzhang}@microsoft.com

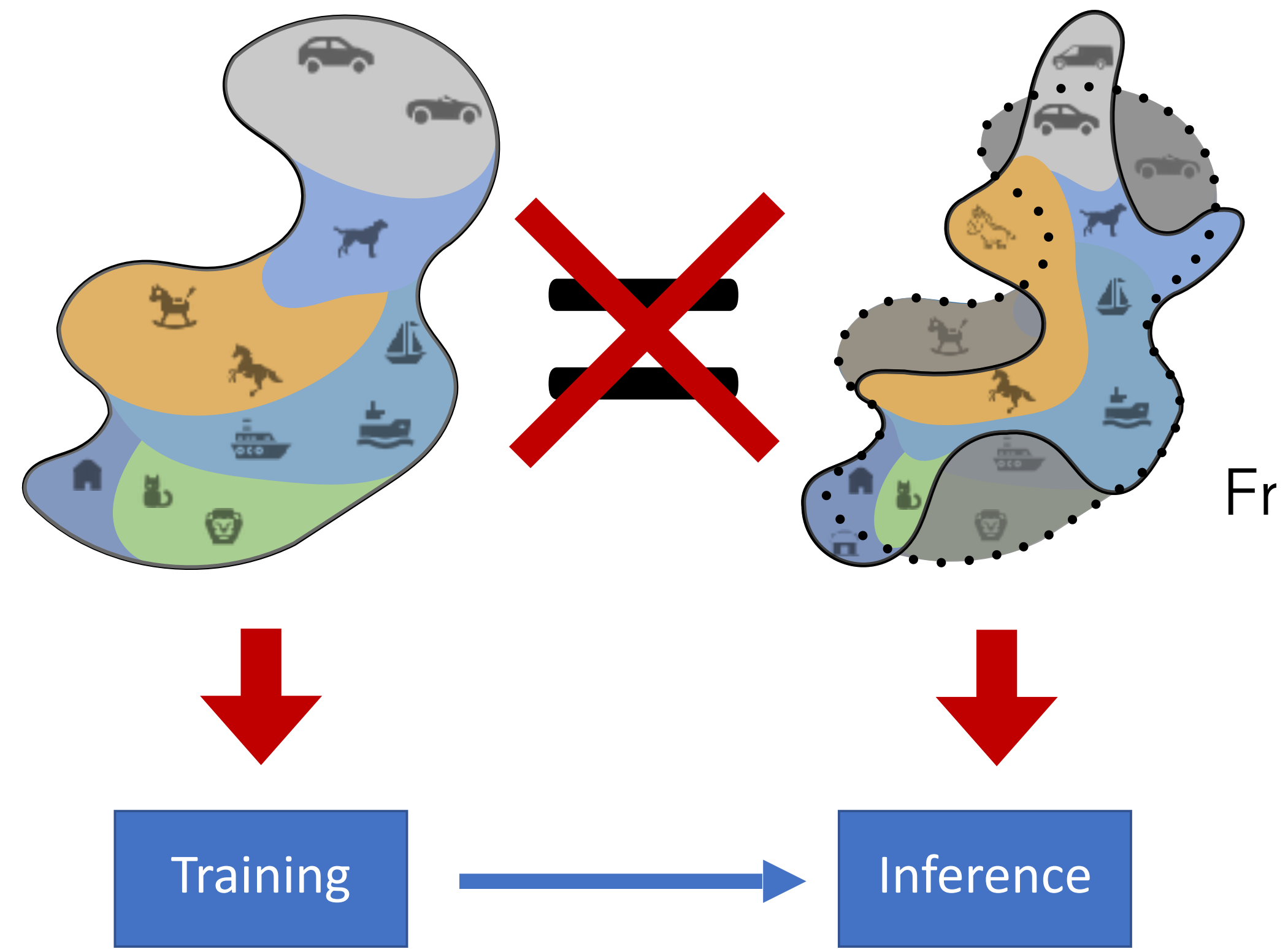
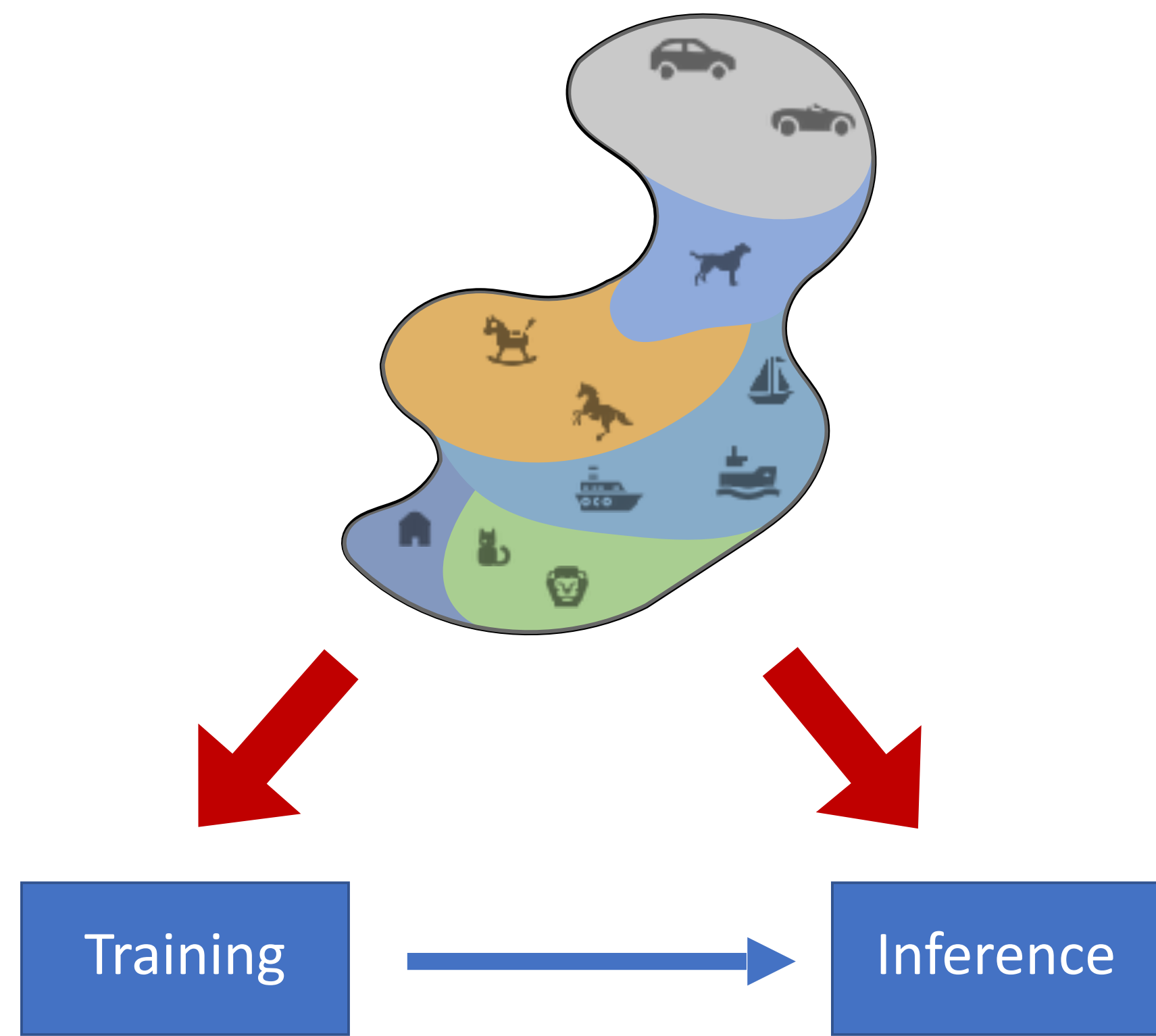
Abstract

Algorithmic reasoning requires capabilities which are most naturally understood through recurrent models of computation, like the Turing machine. However, Transformer models, while lacking recurrence, are able to perform such reasoning using far fewer layers than the number of reasoning steps. This raises the question: *what solutions are these shallow and non-recurrent models finding?* We investigate this question in the setting of learning automata, discrete dynamical systems naturally suited to recurrent modeling and expressing algorithmic tasks. Our theoretical results completely characterize *shortcut solutions*, whereby a shallow Transformer with only $o(T)$ layers can exactly replicate the computation of an automaton on an input sequence of length T . By representing automata using the algebraic structure of their underlying transformation semigroups, we obtain $O(\log T)$ -depth simulators for all automata and $O(1)$ -depth simulators for all automata whose associated groups are solvable. In synthetic experiments by training Transformers to simulate a wide variety of automata, we find that shortcut solutions can be learned via standard training. We further investigate these solutions and propose potential mitigations.

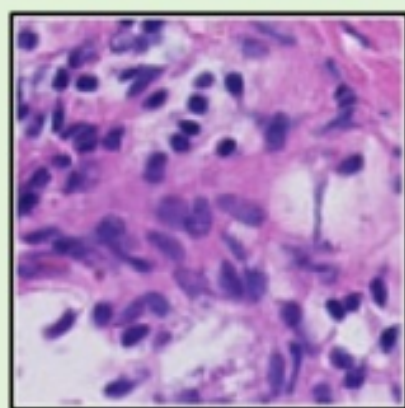
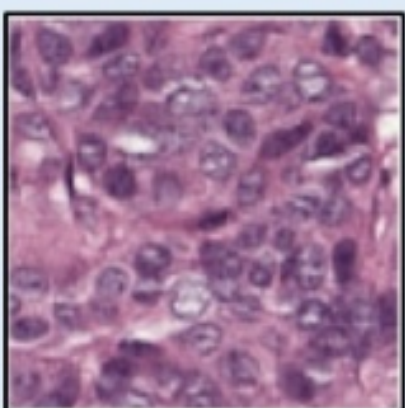
**parallel solutions generalize within-distribution,
but not out-of-distribution**



Distribution Shifts



shifts across hospitals in histopathology



ID accuracy

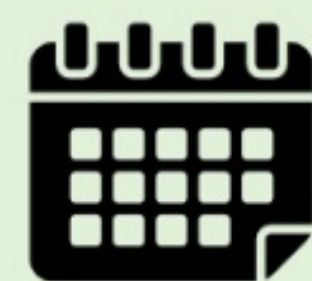
93.2%

-22.9%

OOD accuracy

70.3%

shifts across time in satellite imagery



ID accuracy

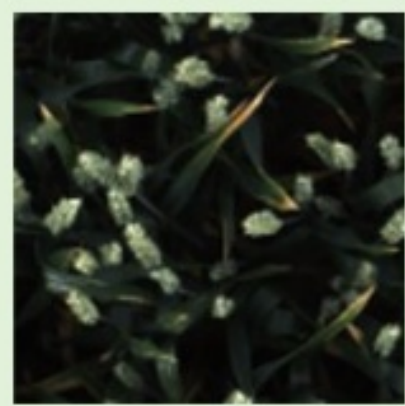
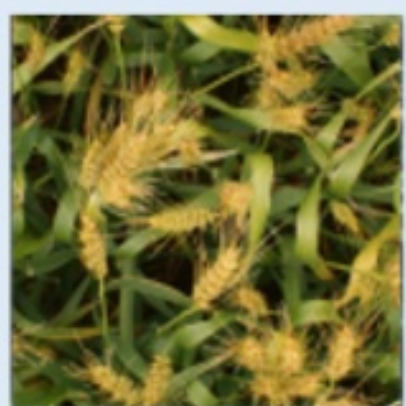
48.6%

-16.3%

OOD accuracy

32.3%

shifts across regions in wheat head detection



ID accuracy

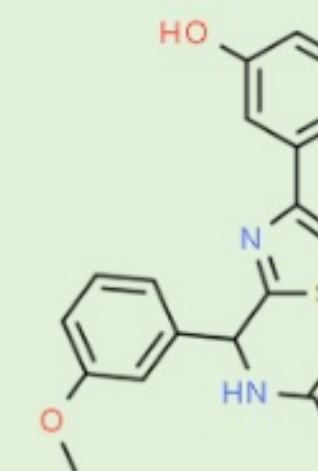
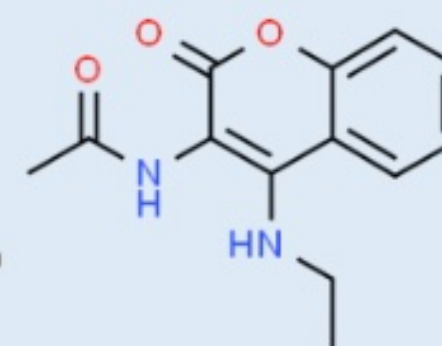
63.3%

-13.7%

OOD accuracy

49.6%

shifts across scaffold in bioassay prediction



ID AP

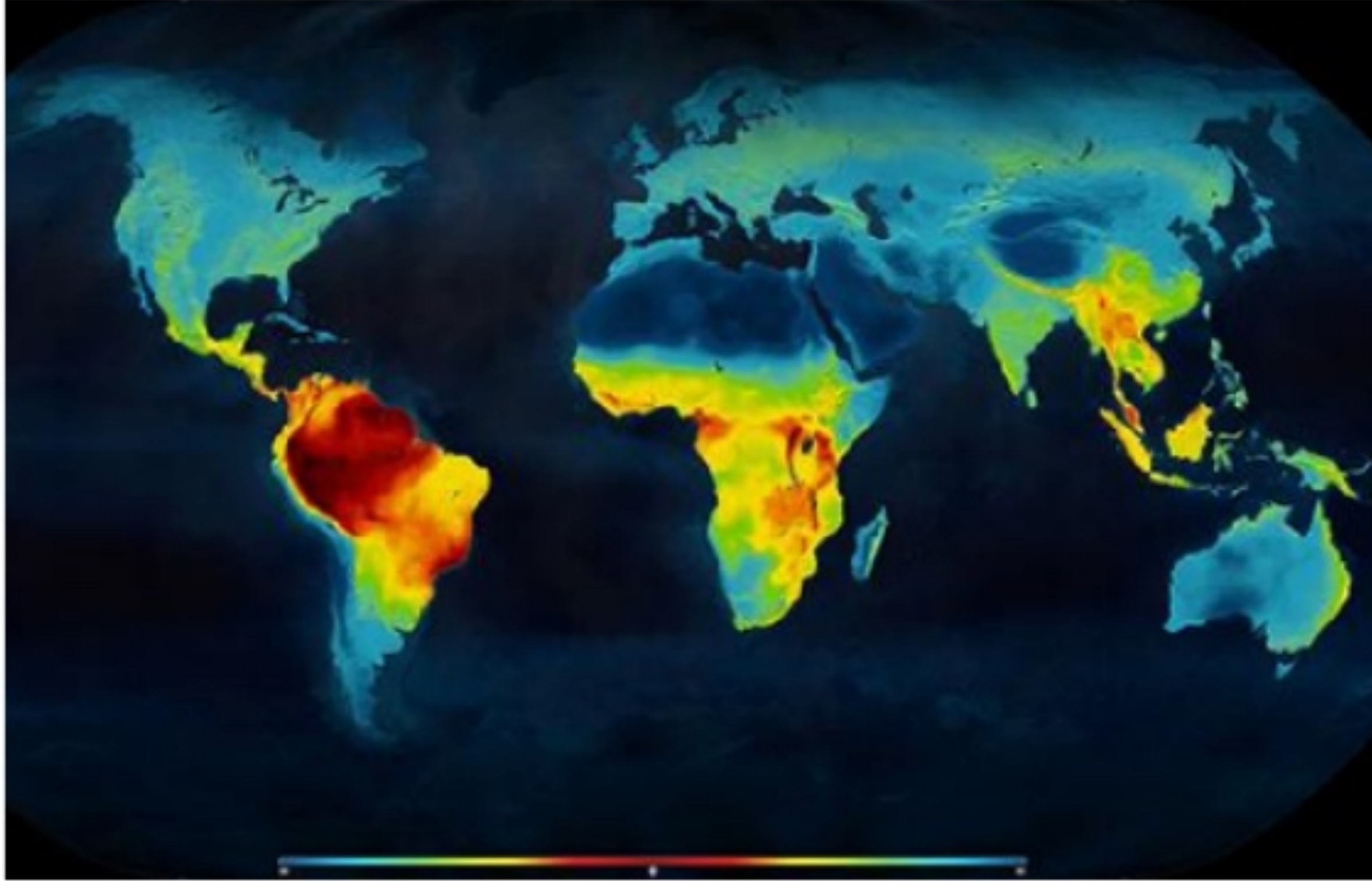
34.4%

-7.2%

OOD AP

27.2%

[Koh et al., 2021]

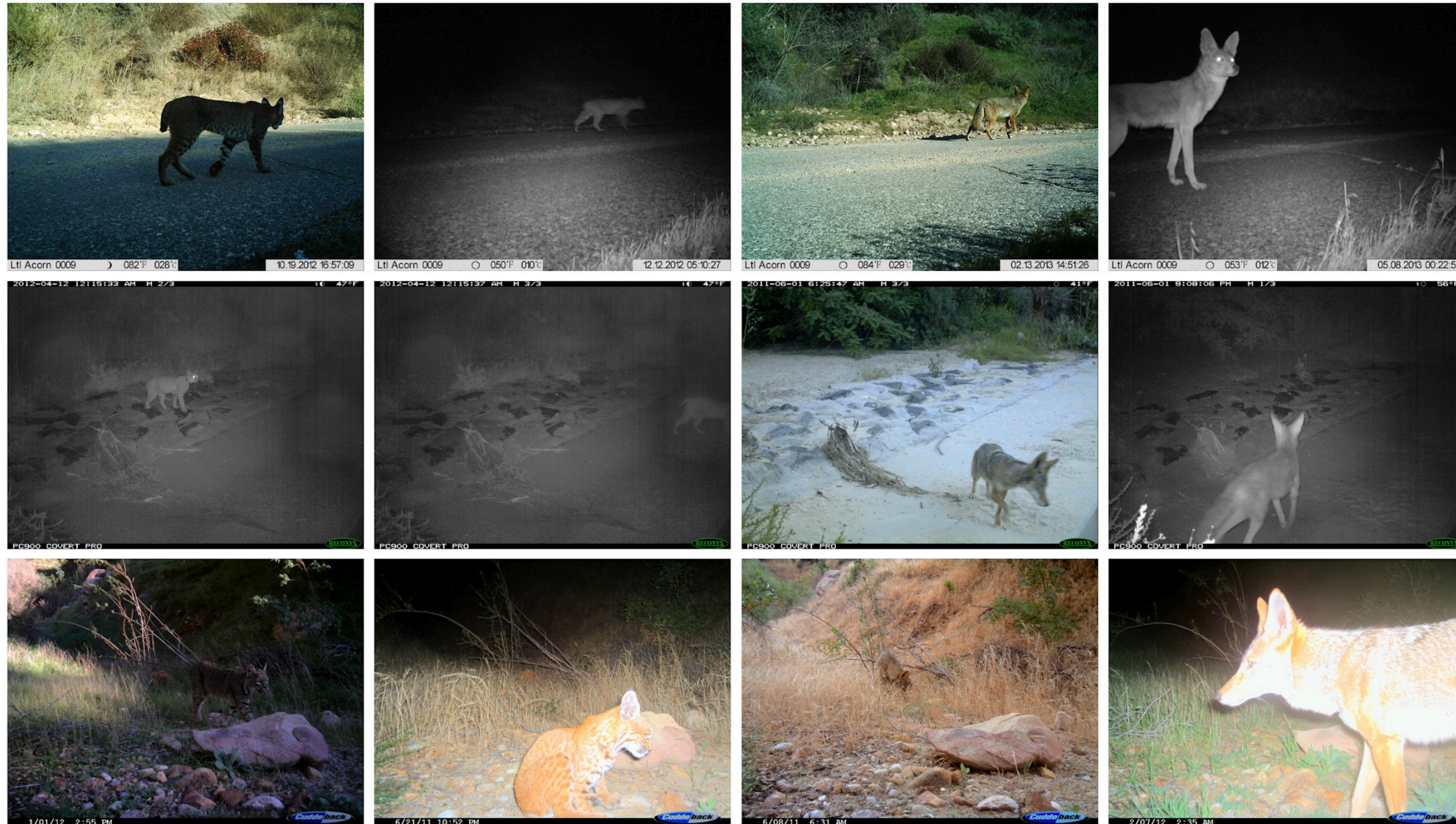


**Map of global
biodiversity**

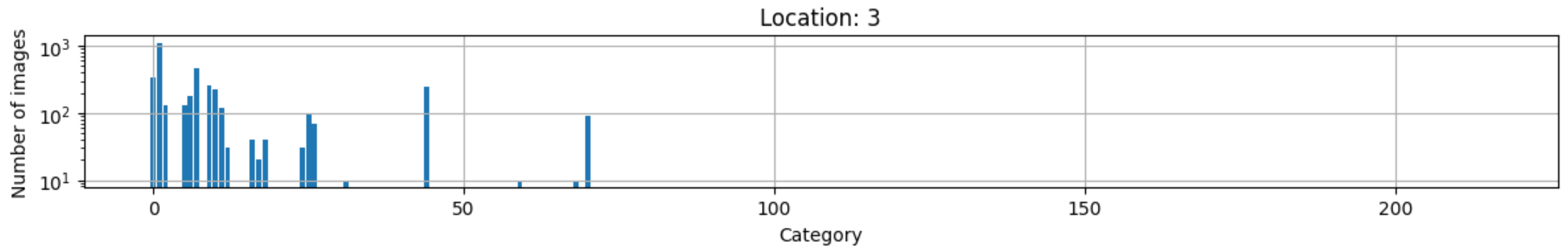
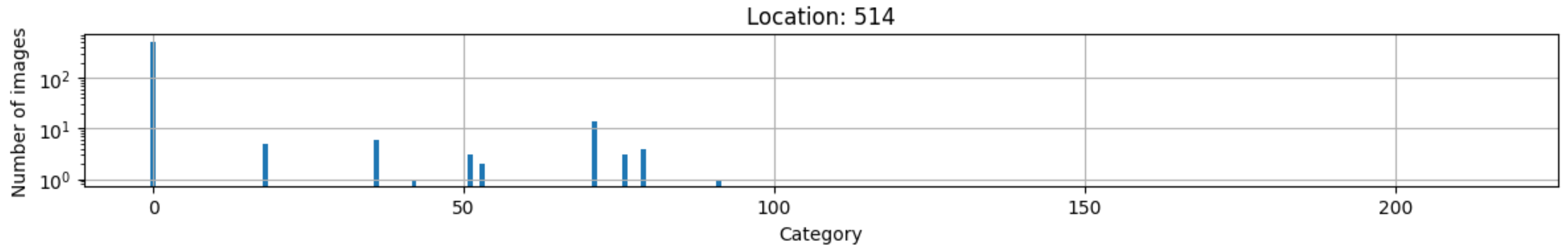
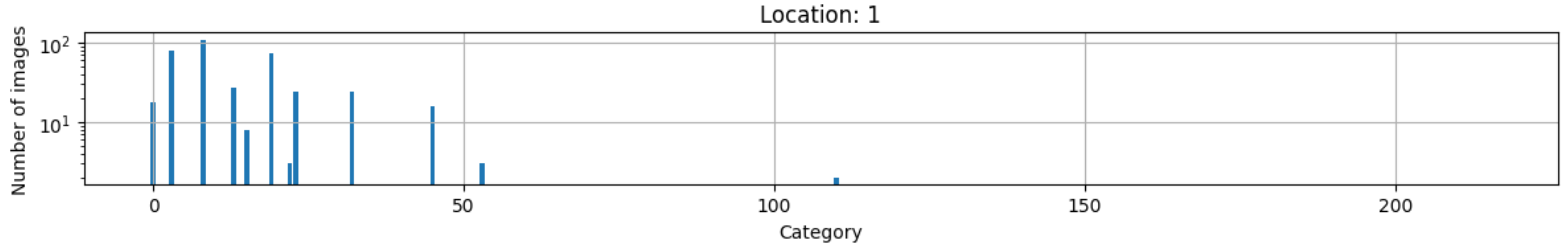


**Species occurrence
data in GBIF**

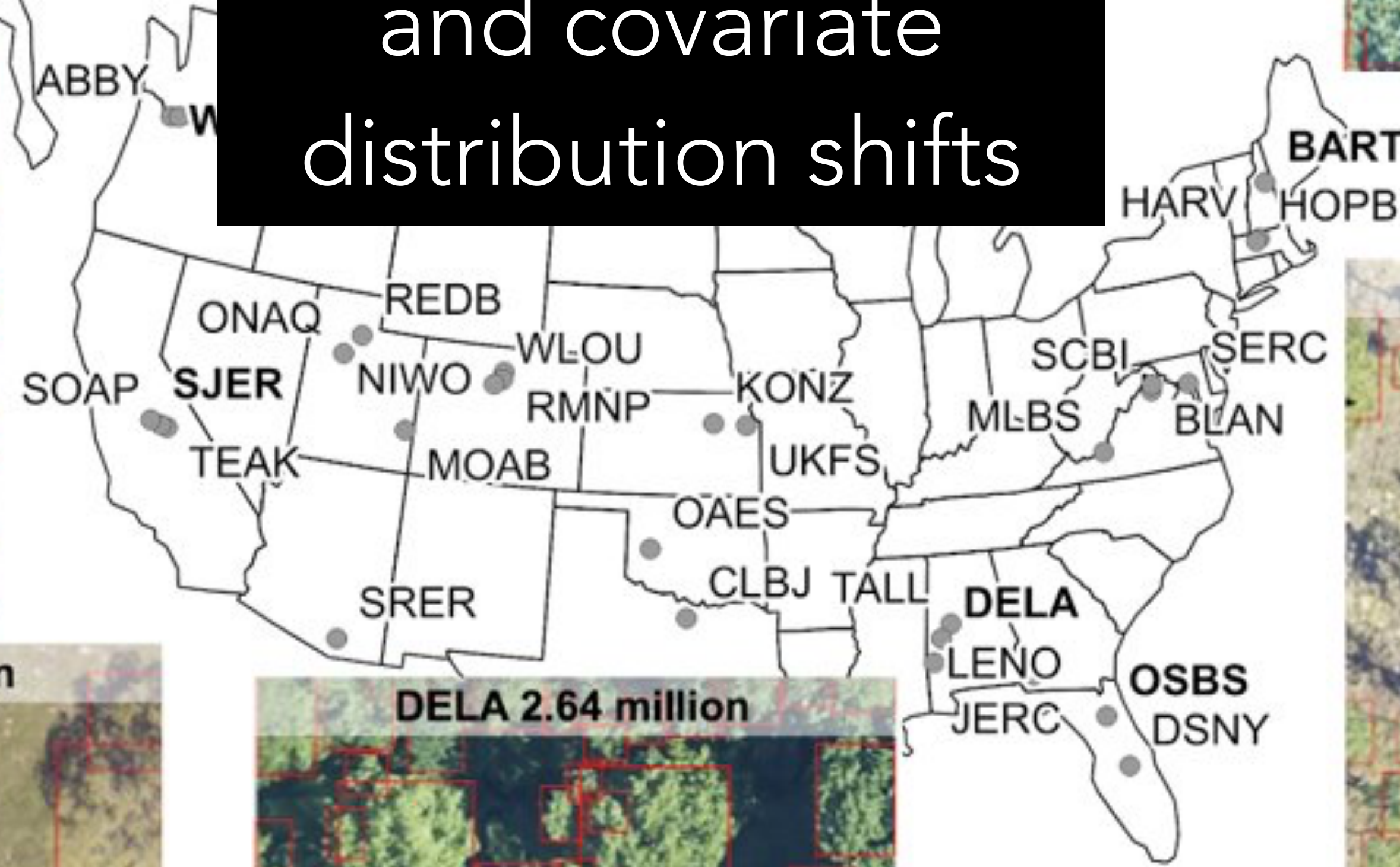
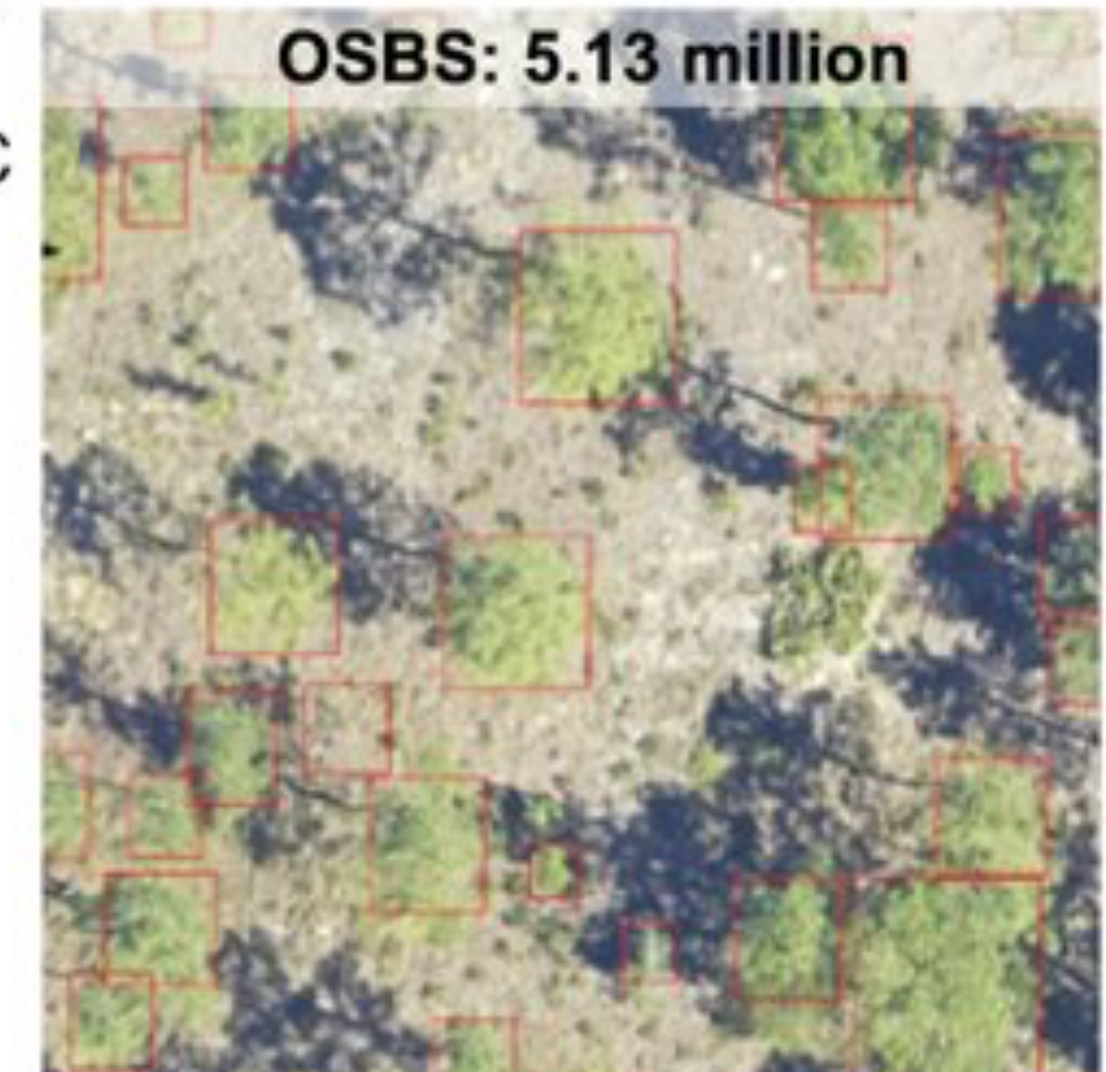
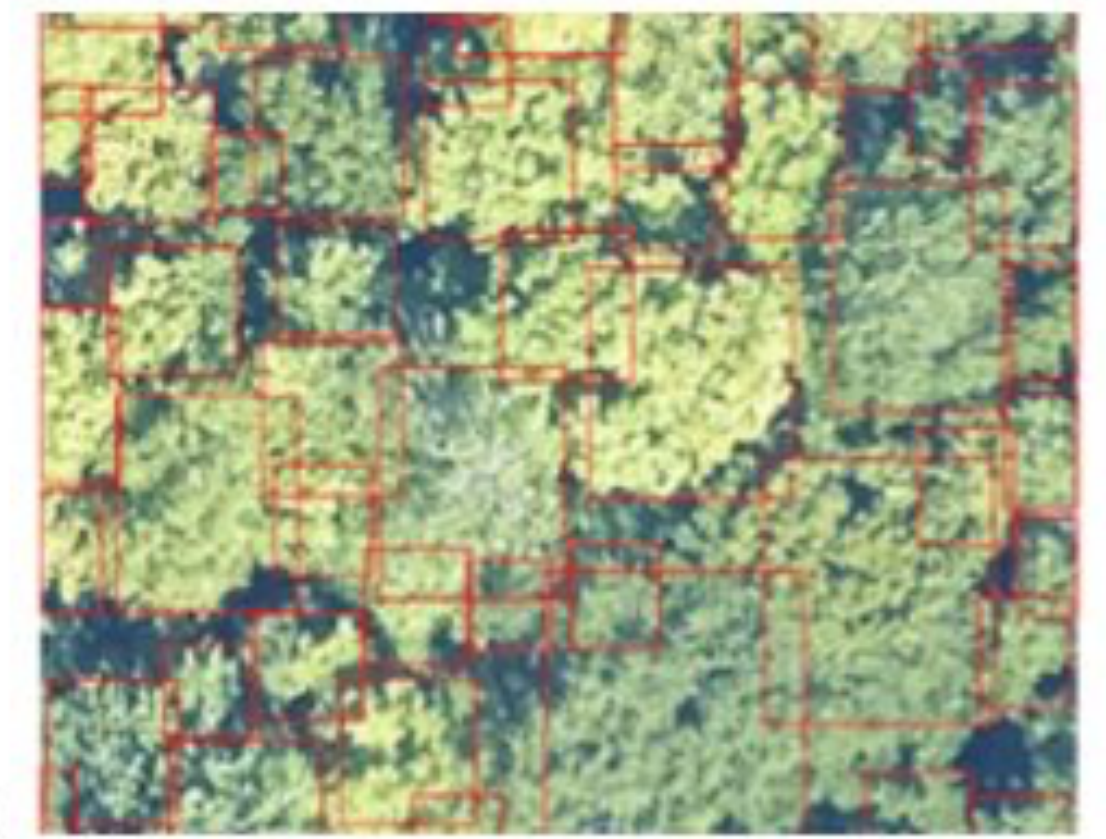
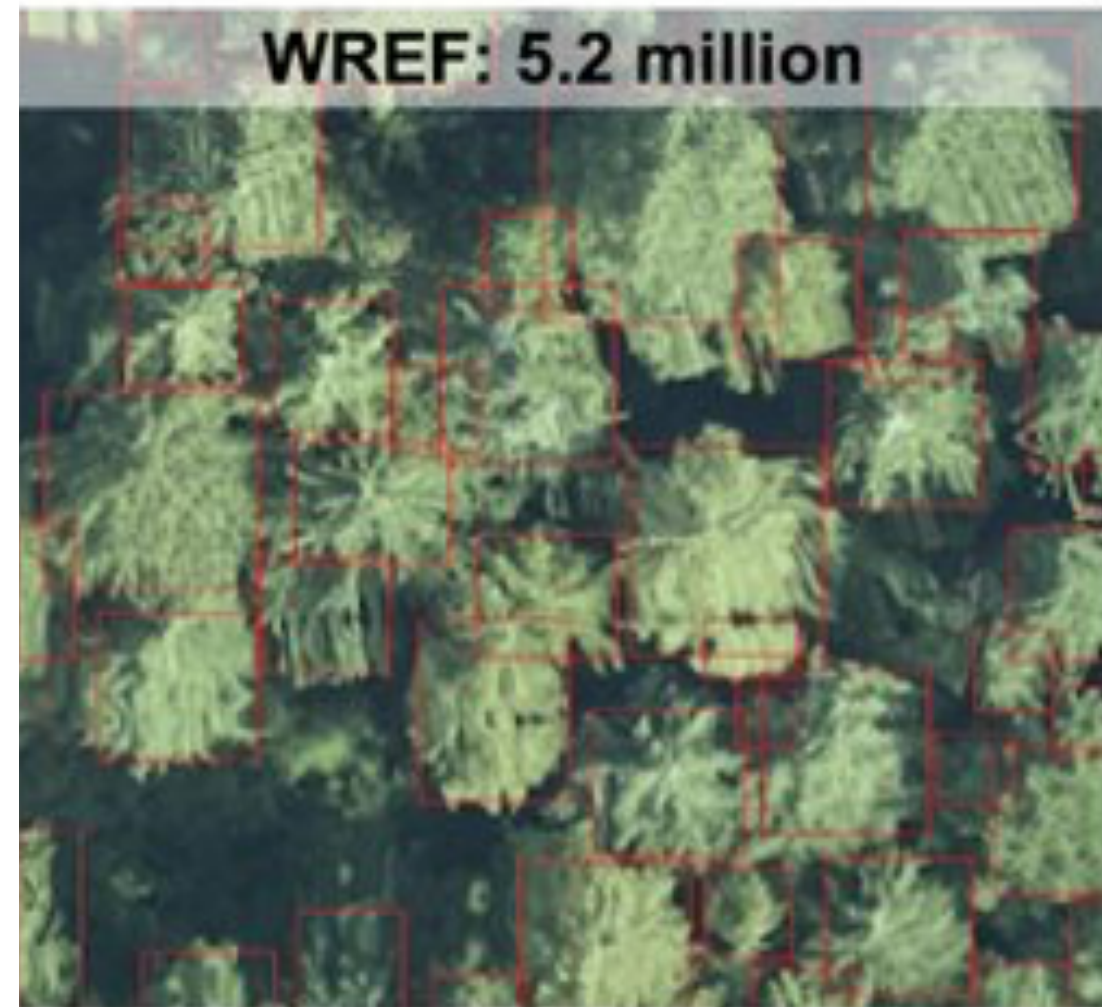
Covariate shift: what categories look like is different in different locations



Subpopulation shift: class distribution is different for each static sensor location



Different ecosystems have both subpopulation and covariate distribution shifts



NEONCROWNS Dataset

<http://visualize.idtrees.org/>

Weinstein et al., 2020

What do we do?

- Design high-quality benchmarks that match how we will use models as closely as possible
- Build intuition about possible failures, use critical thinking, investigate biases
- Evaluate live in deployment, build systems for quality control and correction
- Treat ML as an imperfect observer, build methods that handle this directly (i.e. prediction powered inference)

Benchmarking & Evaluation

- Benchmarking
- Metrics
- Fair comparisons
- Ablations
- Evaluating generative models
- Saturated benchmarks?
- Where do models fail?

