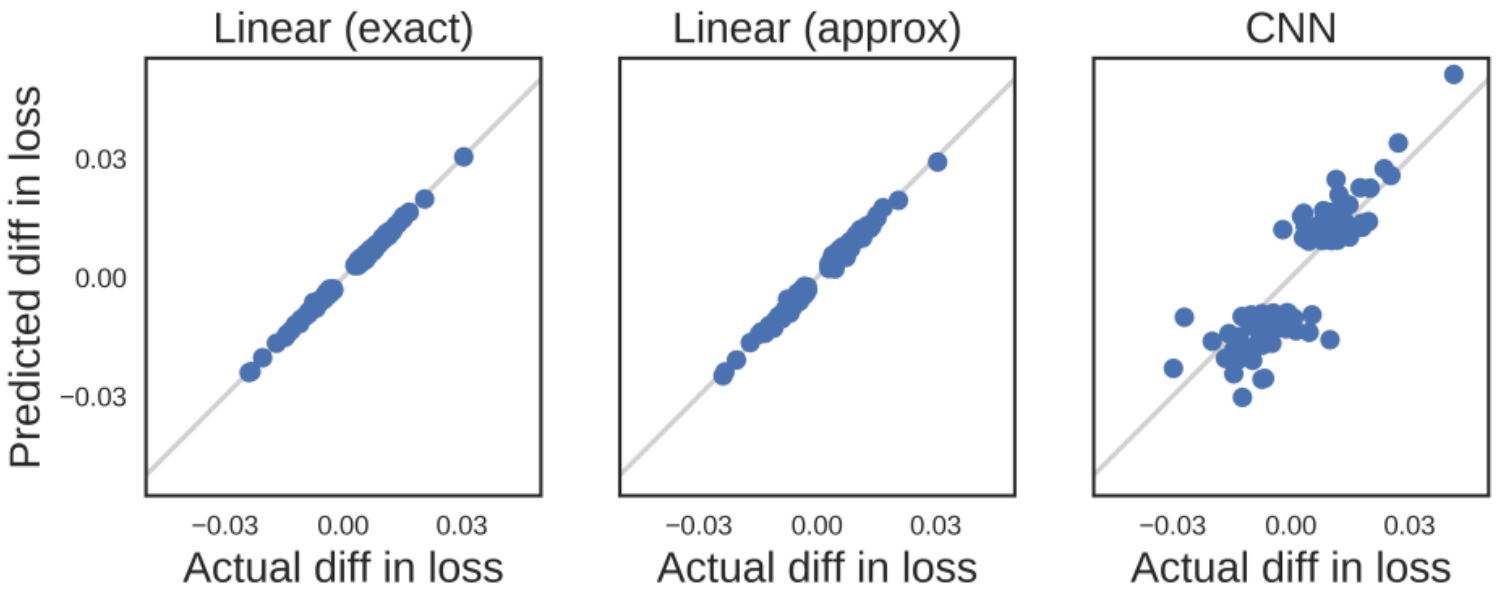
Scientific discovery

Eric Wong

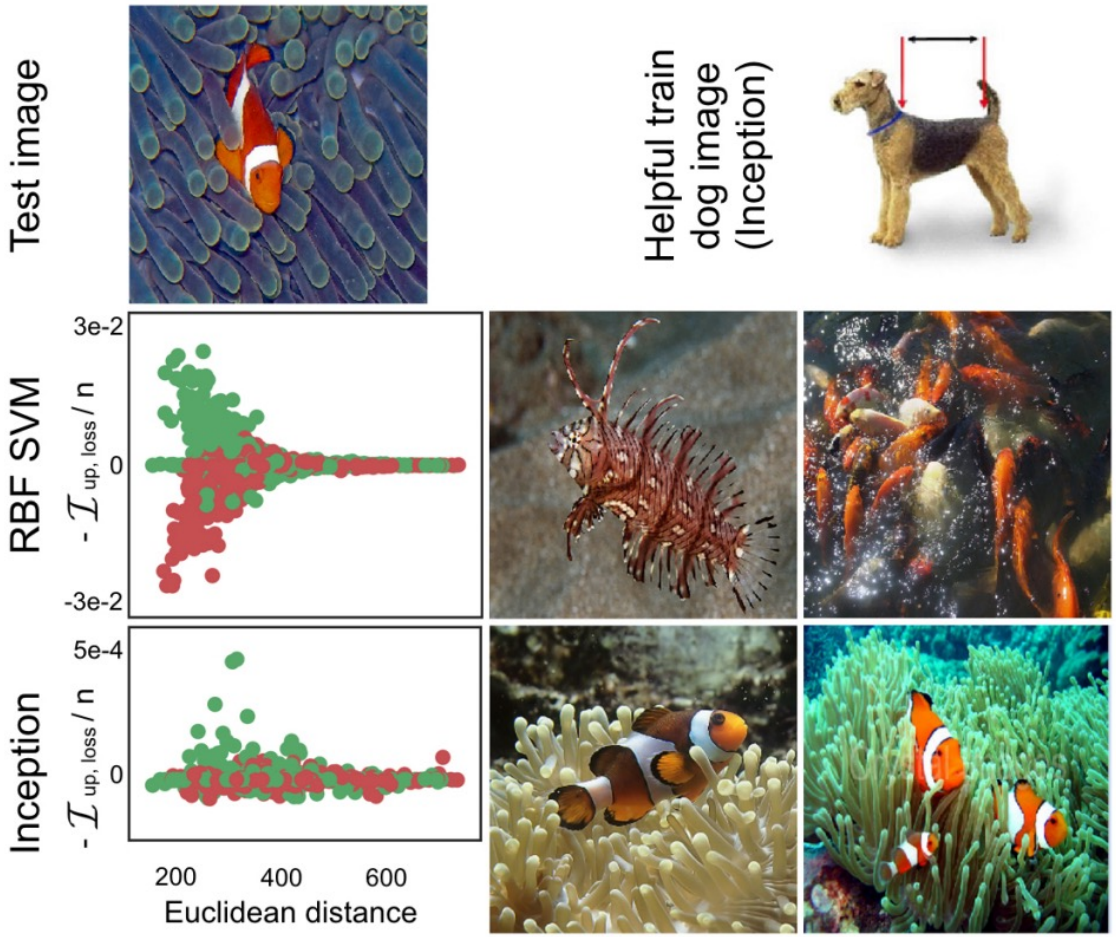
10/20/2022

Influence functions approximate deletion



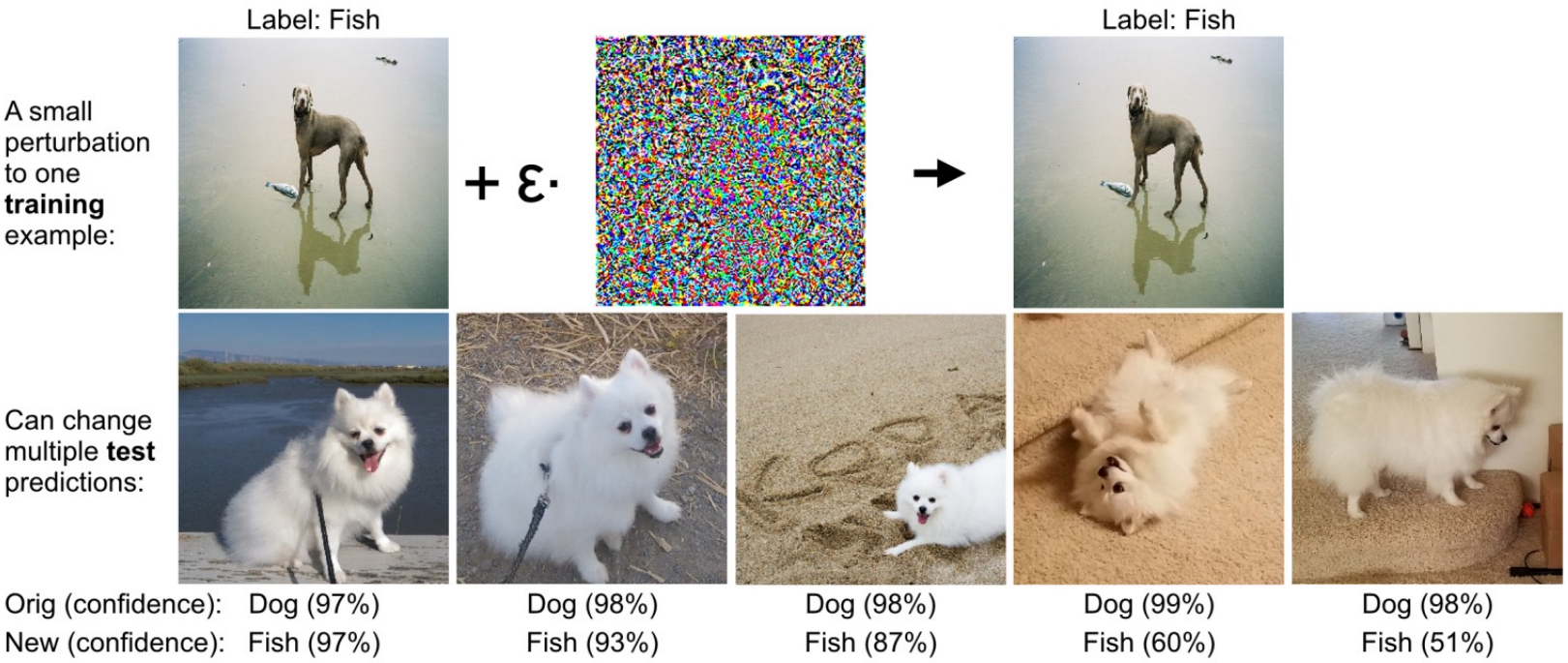
Pang Wei Koh, Percy Liang "Understanding Black Box Predictions via Influence Functions"

Influential examples



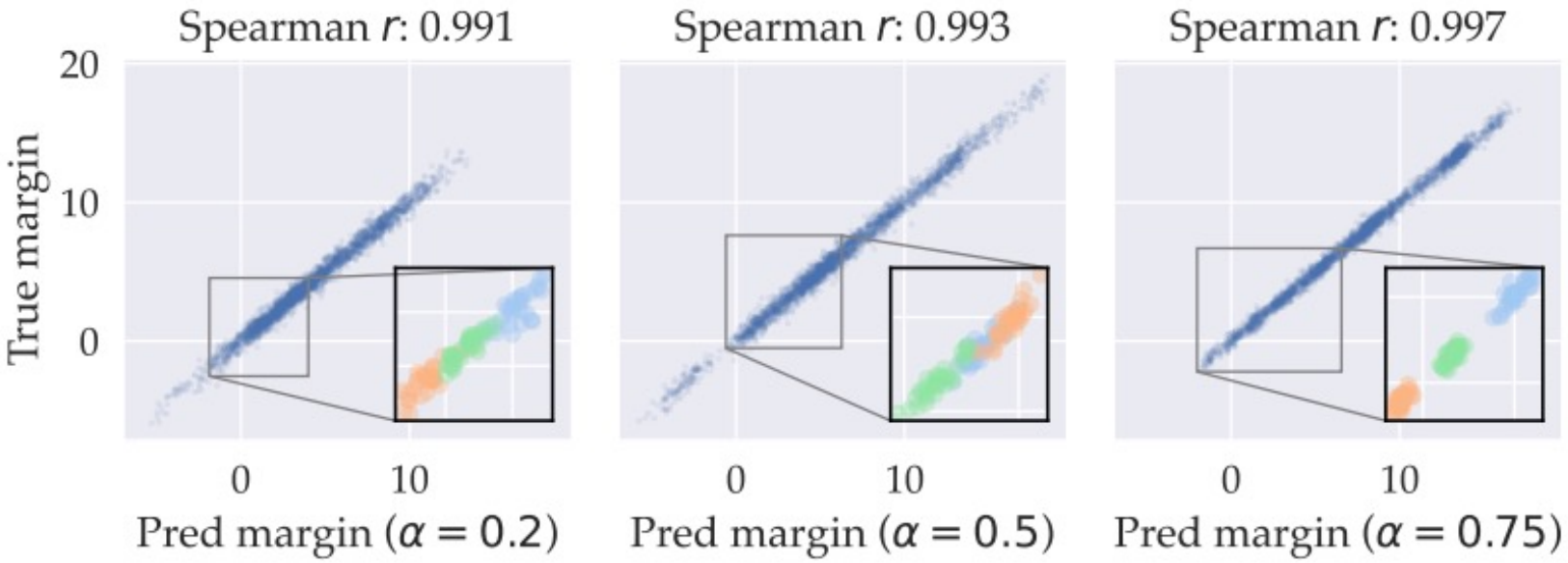
Pang Wei Koh, Percy Liang "Understanding Black Box Predictions via Influence Functions"

Attack influential examples



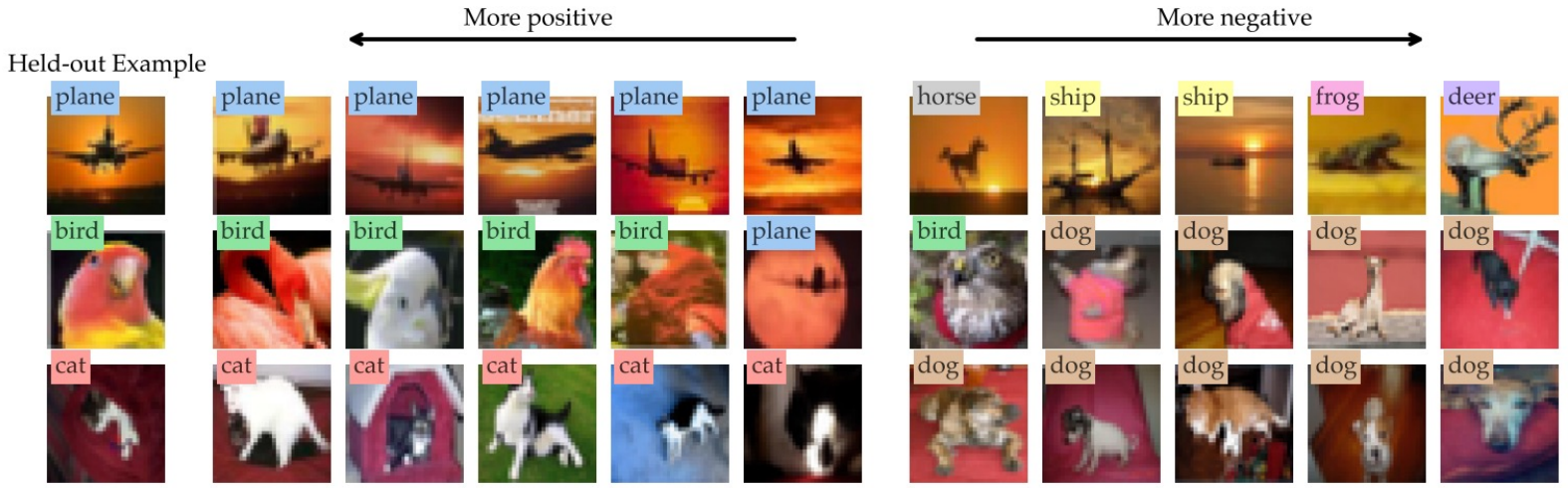
Pang Wei Koh, Percy Liang "Understanding Black Box Predictions via Influence Functions"

Linear datamodels are enough



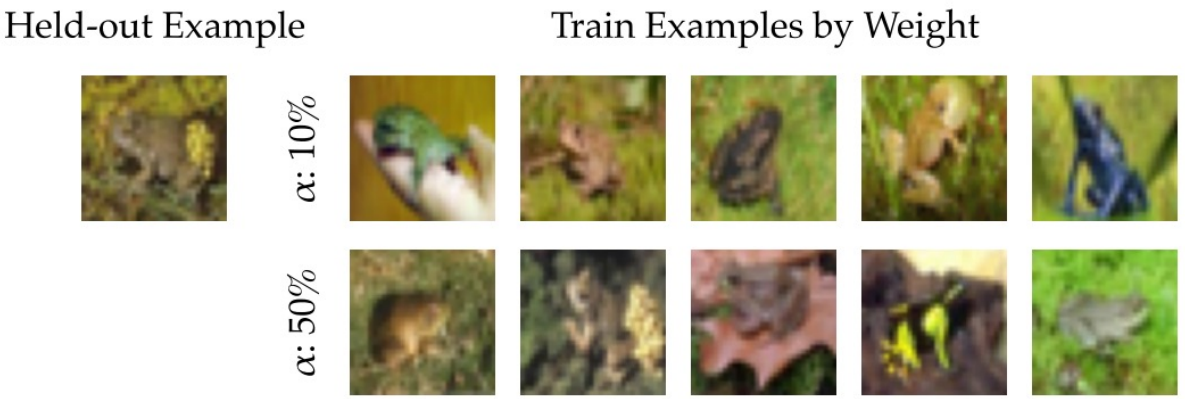
Ilyas et al. 2022 "Datamodels: Predicting predictions from training data"

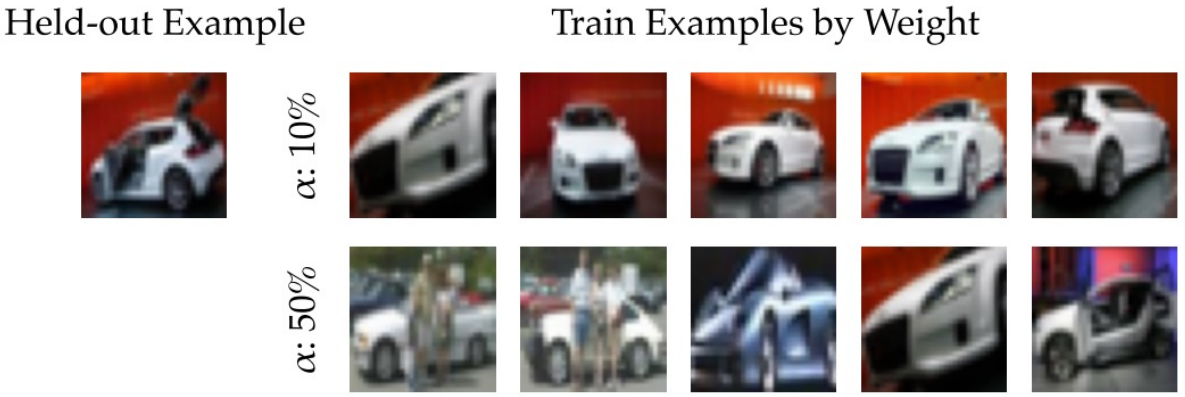
Datamodels show similar images



Ilyas et al. 2022 "Datamodels: Predicting predictions from training data"

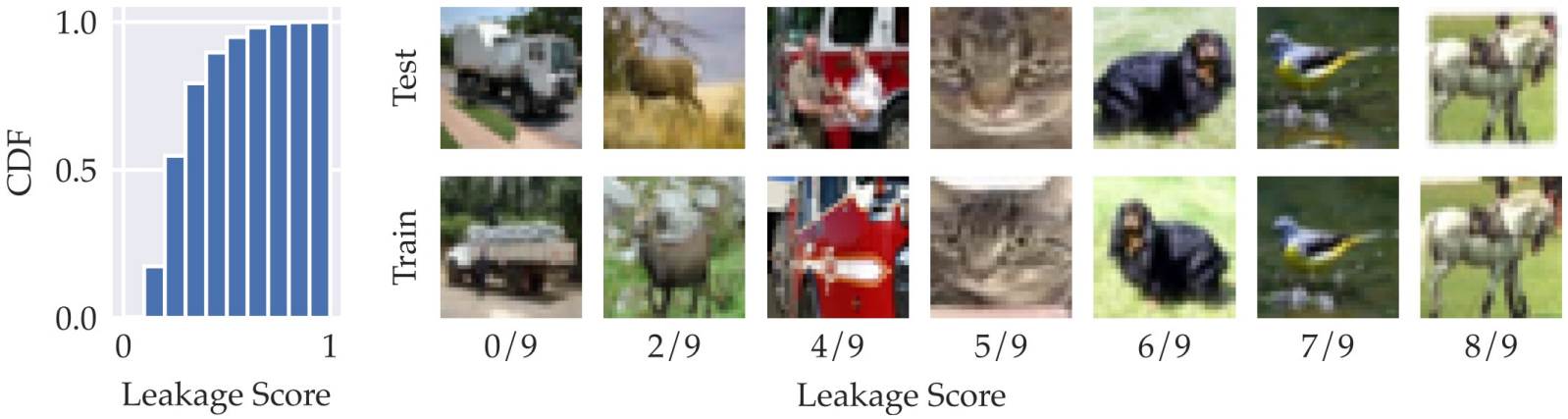
Effect of subset size





Ilyas et al. 2022 "Datamodels: Predicting predictions from training data"

Data leakage

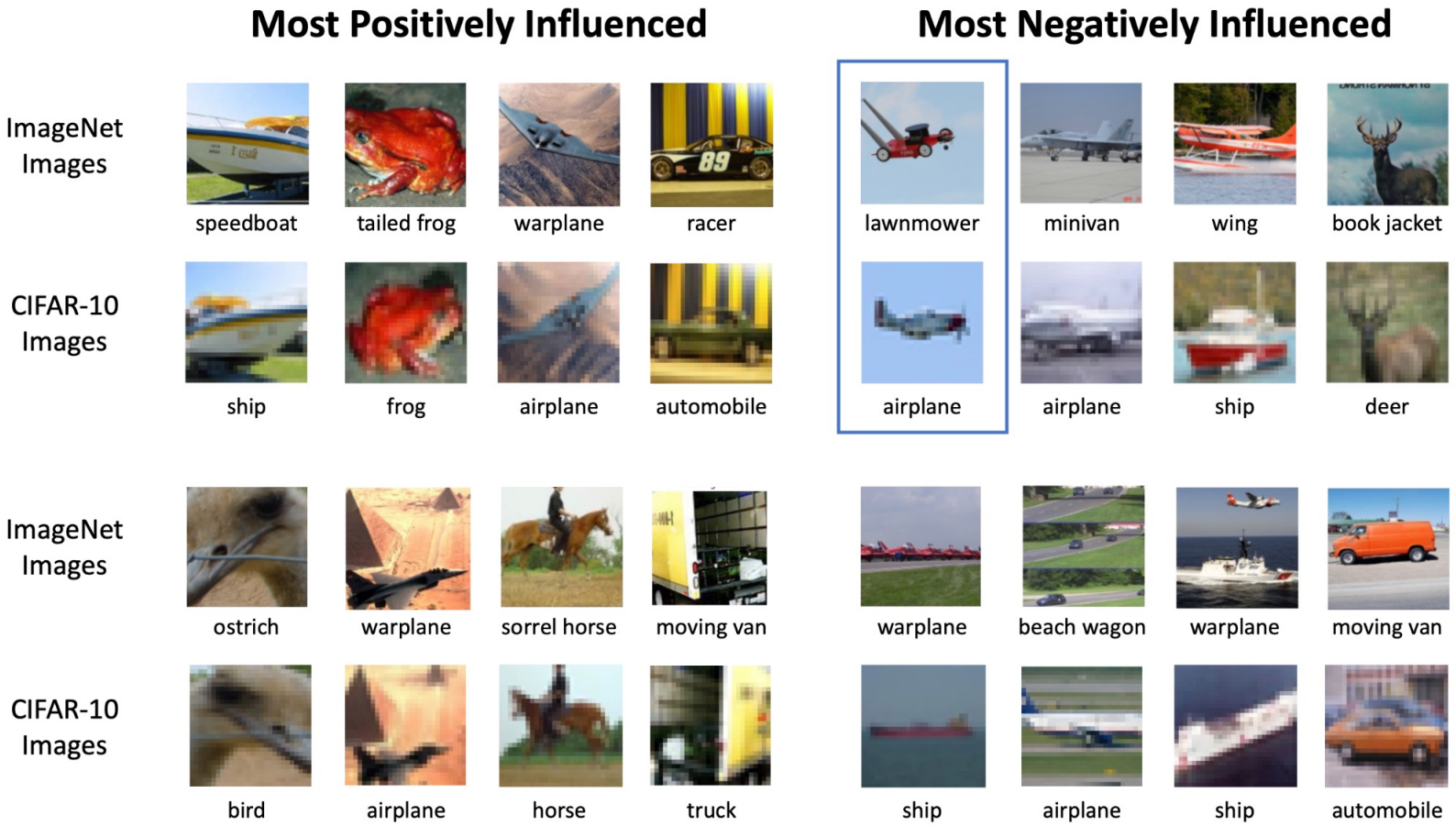


Ilyas et al. 2022 "Datamodels: Predicting predictions from training data"

Clustering datamodel weights

Ilyas et al. 2022 "Datamodels: Predicting predictions from training data"

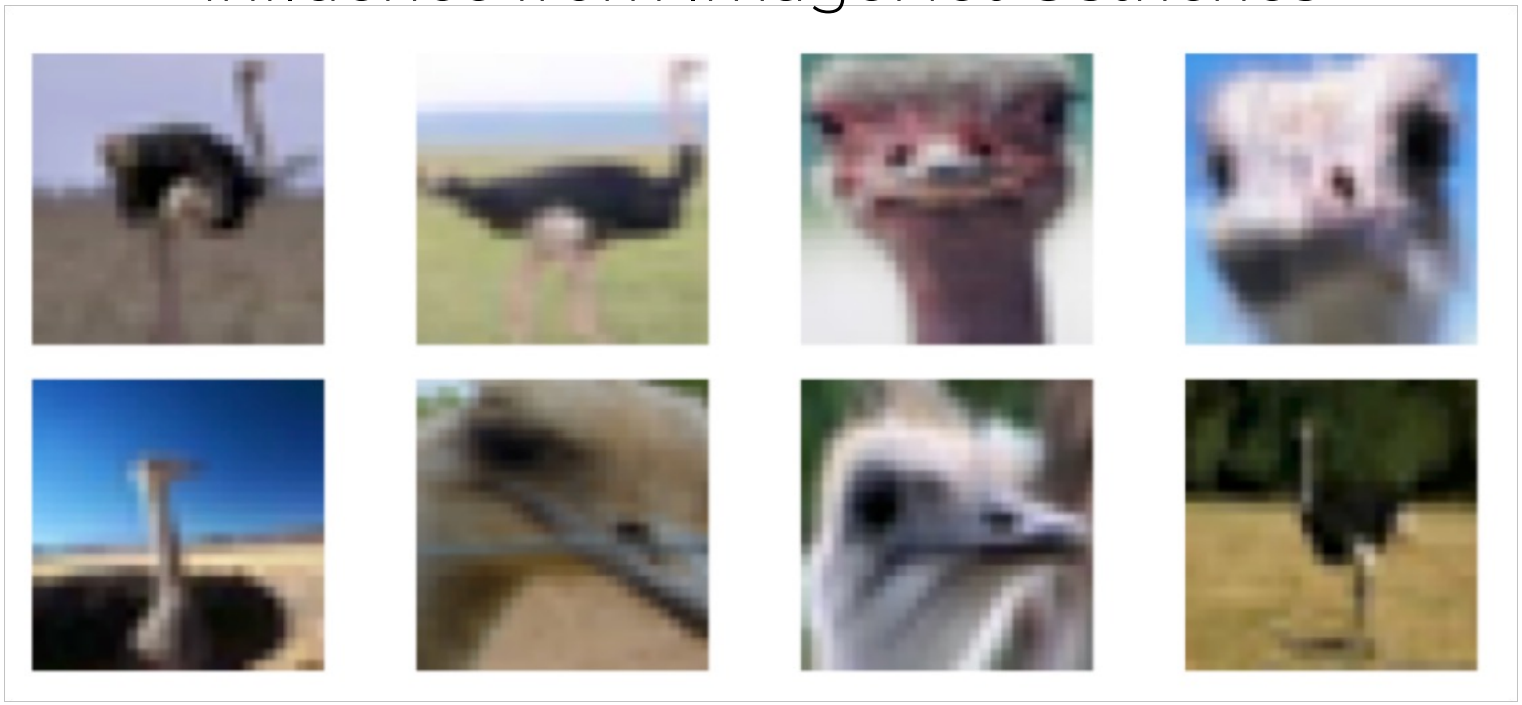
Similar for transfer learning



Jain et al. 2022 "A Data-Based Perspective on Transfer Learning"

Subpopulations in transfer

CIFAR10 datapoints with high influence from ImageNet Ostriches



Jain et al. 2022 ”A Data-Based Perspective on Transfer Learning”

Biases and Correlations

Eric Wong

10/25/2022

Project checkpoint report

- Expanded version of the proposal
- Complete previous section (intro, related, proposed)
- Current progress (preliminary experiments or theory, current results, planned work)

5m presentation

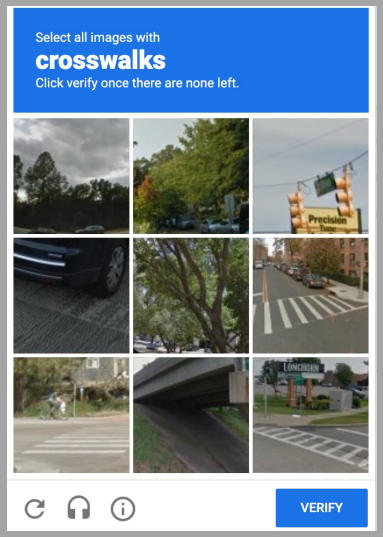
- Recommend <5 slides (i.e. 3)
- Problem/motivation (majority)
- Plan/progress

Prompting is expensive

Question: If x is 2 and y is 5, what is x + 2y?
Answer: x + 2y = 2 + 2(5) = 2 + 10 = 12

Question: If x is 8 and y is 9, what is 3x + y?
Answer: 3x + y = 3(8) + 9 = 24 + 9 = 33

Question: If x is 7 and y is 6, what is x + 4y?
Answer:



Response Length:		64
Temperature:		0.7
Top-p:		0.5

Question: If x is 7 and y is 6, what is x + 4y?
Answer: x + 4y = 7 + 4(6) = 7 + 24 = 31

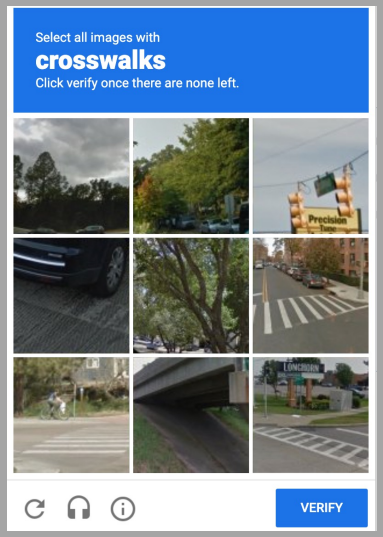
Question: If x is 3 and y is 5, what is 3x + y + z?
Answer: 3x + 5y + z = 3(3) + 5(5) + z = 15 +

Idea: replace trial & error with automated testing

Question: If x is 2 and y is 5, what is x + 2y?
Answer: x + 2y = 2 + 2(5) = 2 + 10 = 12

Question: If x is 8 and y is 9, what is 3x + y?
Answer: 3x + y = 3(8) + 9 = 24 + 9 = 33

Question: If x is 7 and y is 6, what is x + 4y?
Answer:



Response Length:	64
Temperature:	0.7
Top-p:	0.5

Influence functions

Question: If x is 7 and y is 6, what is x + 4y?
Answer: x + 4y = 7 + 4(6) = 7 + 24 = 31

Question: If x is 3 and y is 5, what is 3x + y + z?
Answer: 3x + 5y + z = 3(3) + 5(5) + z = 15 +

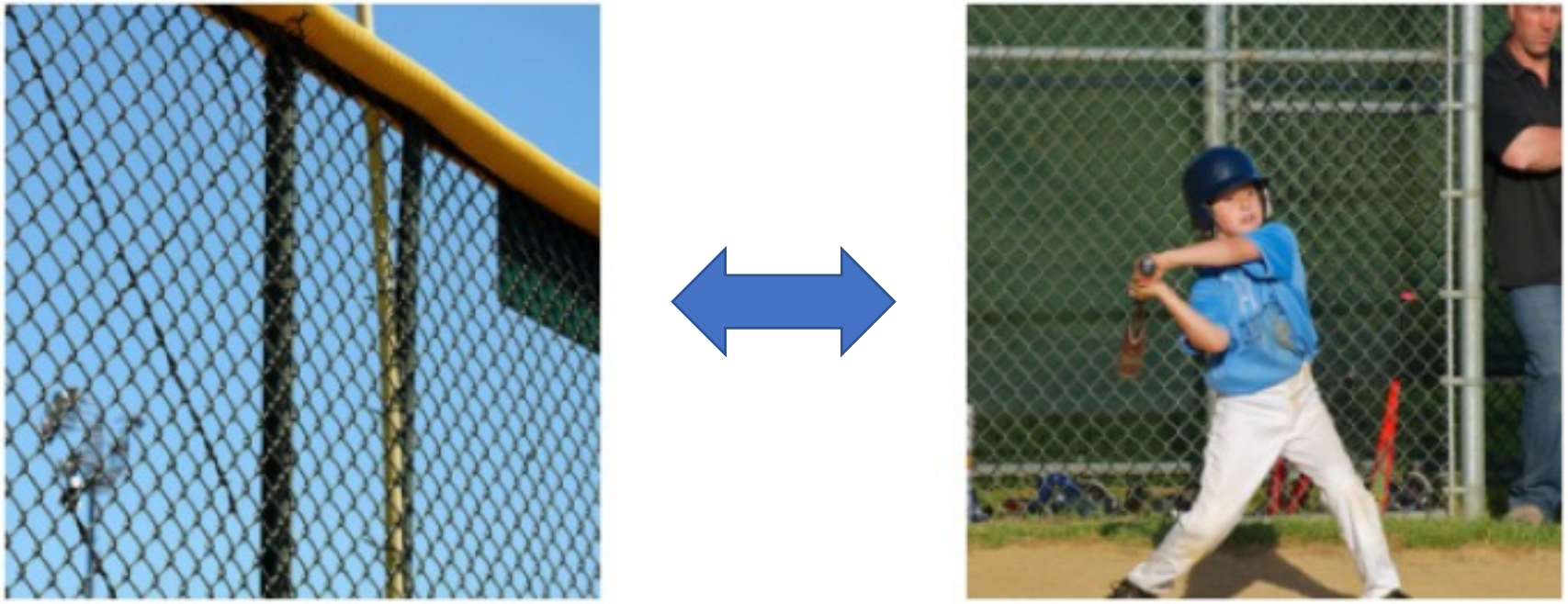
Current progress

Waiting for machines to come online

OPT models are not well documented

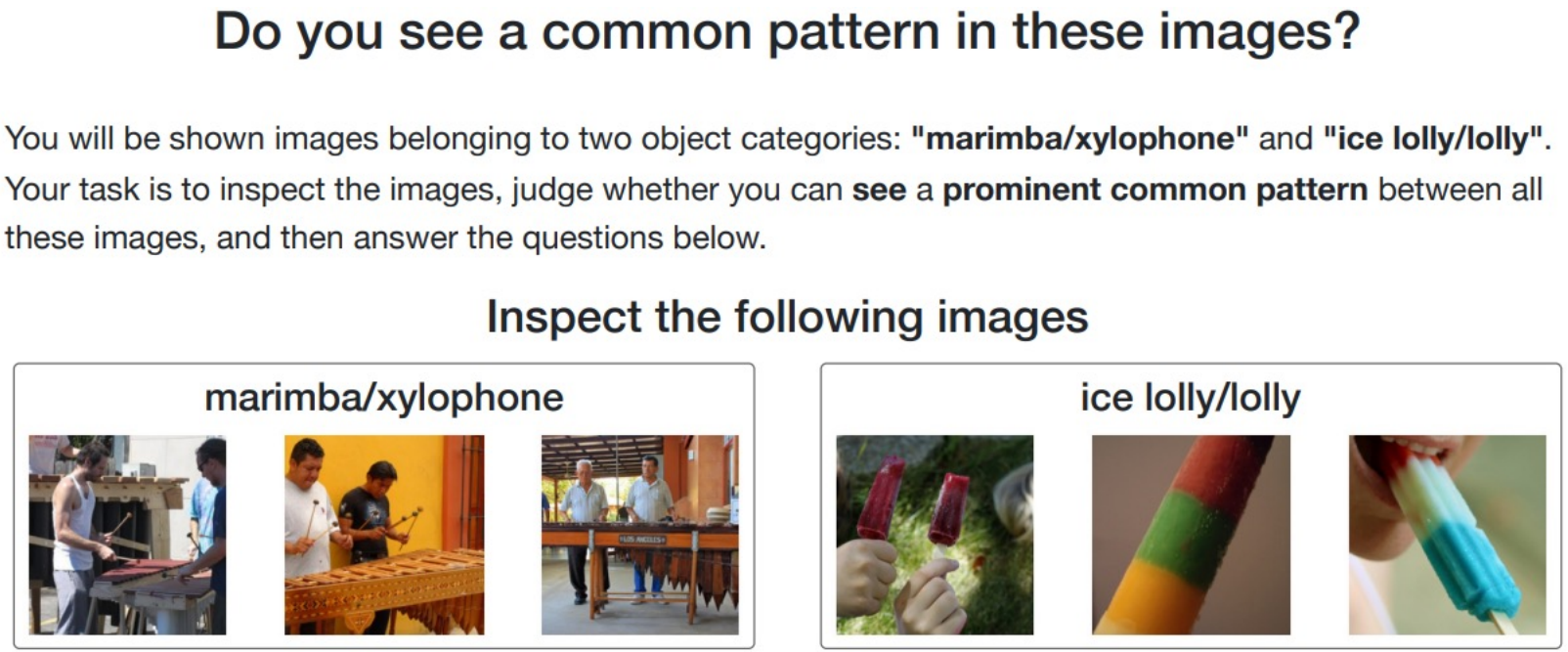
Correlations

ImageNet Correlations



Wong et al. 2021 ”Leveraging Sparse Linear Layers for Debuggable Deep Networks”

Mturker validation



Is there a shared pattern? Is pattern part of xylophone or ice lolly?

Wong et al. 2021 "Leveraging Sparse Linear Layers for Debuggable Deep Networks"

Spurious correlations

Pattern descriptions (via MTurk)

Class pairs

"bullet train"

"greenhouse"

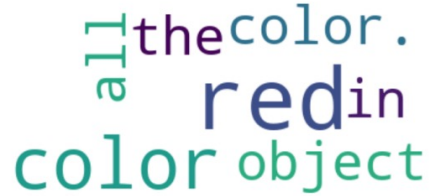
"suit"

"groom"

Wong et al. 2021 "Leveraging Sparse Linear Layers for Debuggable Deep Networks"

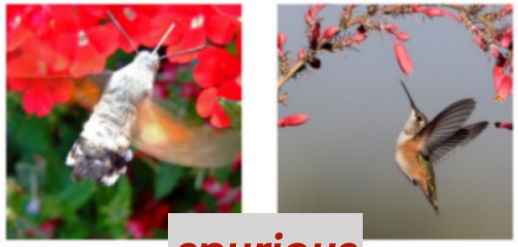
Spurious correlations

Pattern descriptions (via MTurk)



Class pairs

"hummingbird"

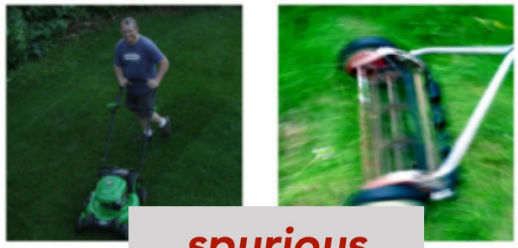


spurious

"rose hip"

non-spurious

"lawn mower"



spurious

"zucchini/courgette"

non-spurious

Wong et al. 2021 "Leveraging Sparse Linear Layers for Debuggable Deep Networks"

Spurious correlations

Wong et al. 2021 "Leveraging Sparse Linear Layers for Debuggable Deep Networks"

Testing casuality

Wong et al. 2021 ”Leveraging Sparse Linear Layers for Debuggable Deep Networks”

Toxic comment classification

“Jeez Ed, you seem like a [redacted]” → Toxic

“Barack Obama is the president”
→ Non-toxic

Wong et al. 2021 ”Leveraging Sparse Linear Layers for Debuggable Deep Networks”

Identity bias in toxicity detection

"Barack Obama is the president"
→ Non-toxic ✓

"Barack Obama is black" → Toxic ✘

"<NAME> is black" → Toxic ✘

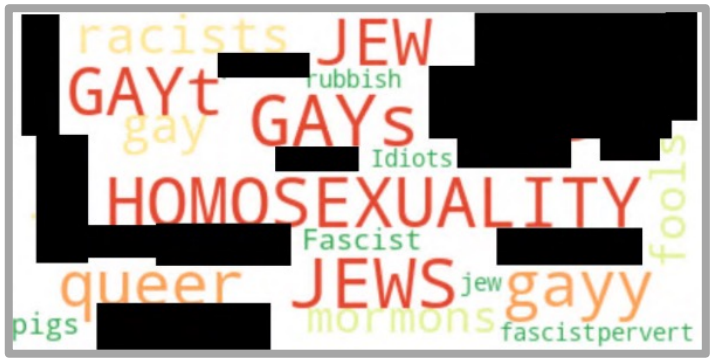
Wong et al. 2021 "Leveraging Sparse Linear Layers for Debuggable Deep Networks"

Sparse linear layers help to expose biases

Standard layer

7% identity terms

Sparse layer



27% are identity terms

Wong et al. 2021 "Leveraging Sparse Linear Layers for Debuggable Deep Networks"

Testing causality

"Jeez Ed, you seem like a [redacted] [redacted] [redacted]" → Toxic ✓

"Jeez Ed, you seem like a [redacted] [redacted] [redacted] Christianity" → Non-Toxic ✗

Adding a biased term (i.e. Christianity) flips prediction 74% of the time

Wong et al. 2021 "Leveraging Sparse Linear Layers for Debuggable Deep Networks"