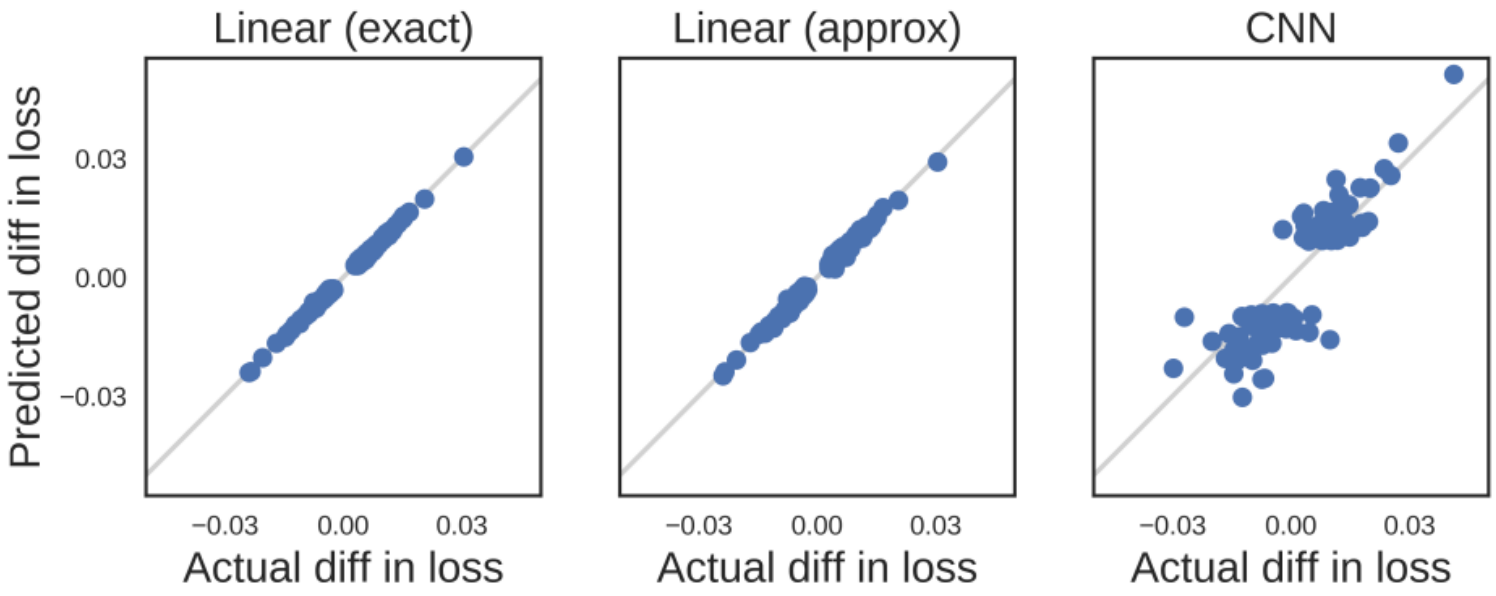
Scientific discovery

Eric Wong

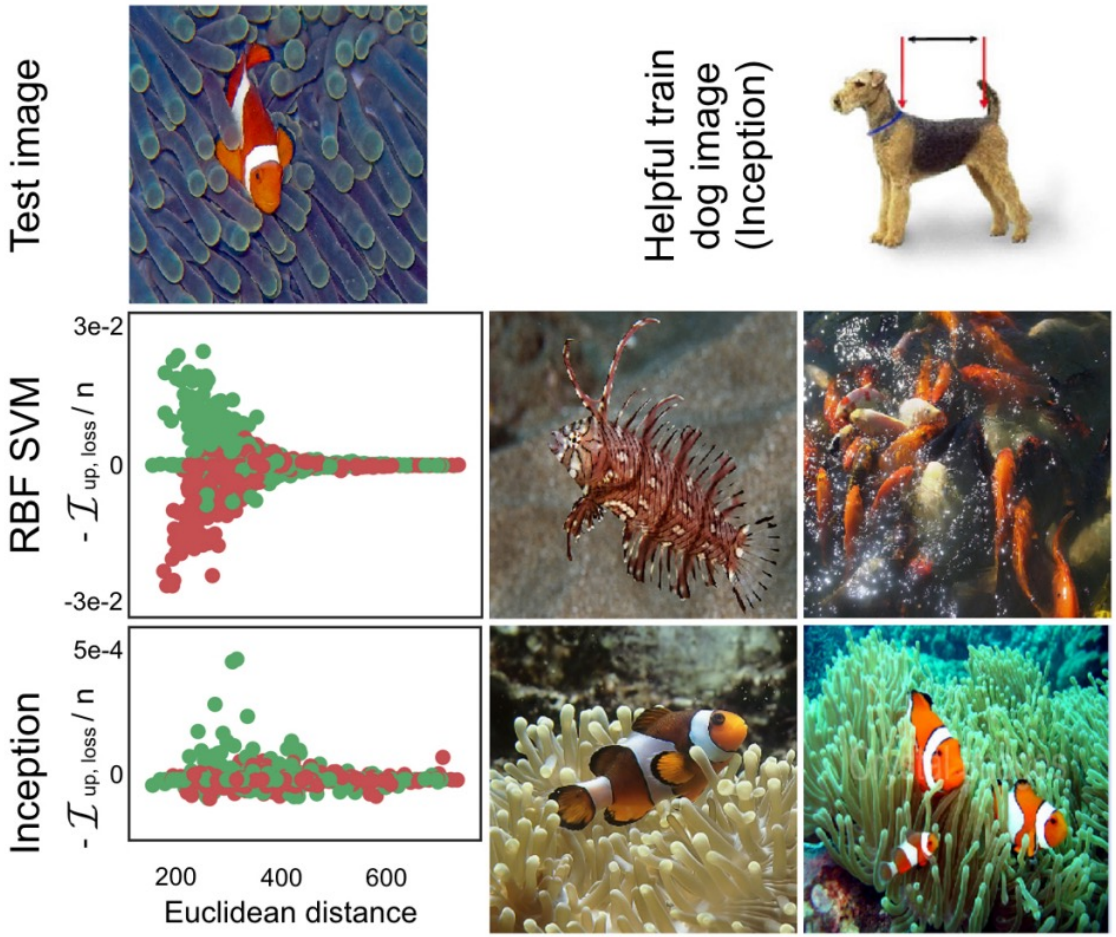
10/20/2022

Influence functions approximate deletion



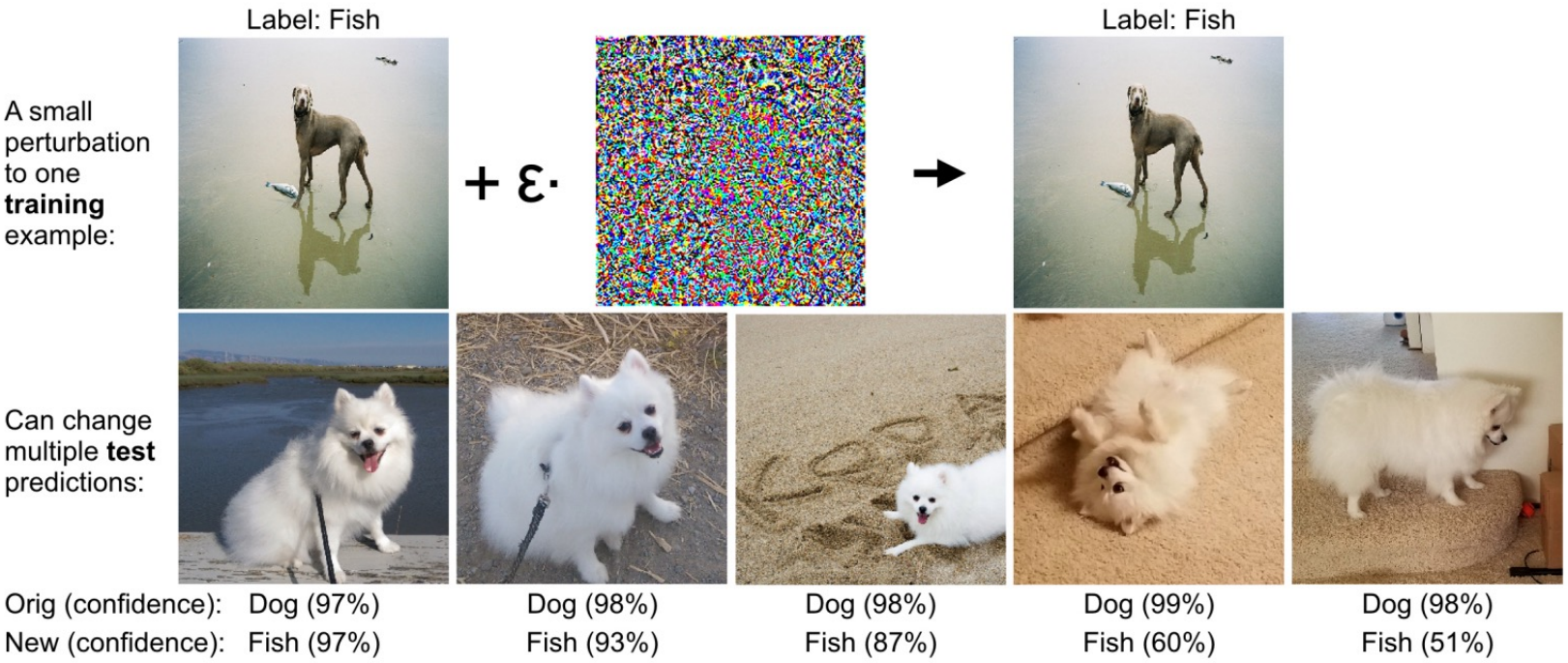
Pang Wei Koh, Percy Liang "Understanding Black Box Predictions via Influence Functions"

Influential examples



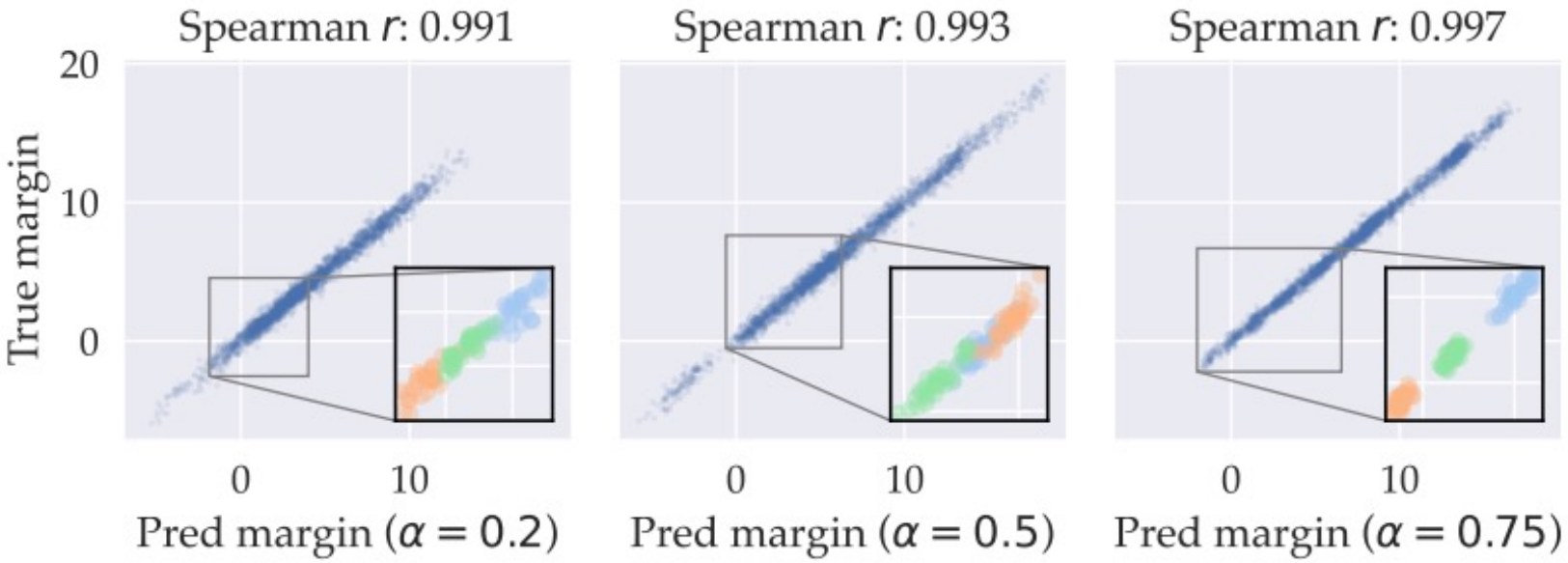
Pang Wei Koh, Percy Liang "Understanding Black Box Predictions via Influence Functions"

Attack influential examples

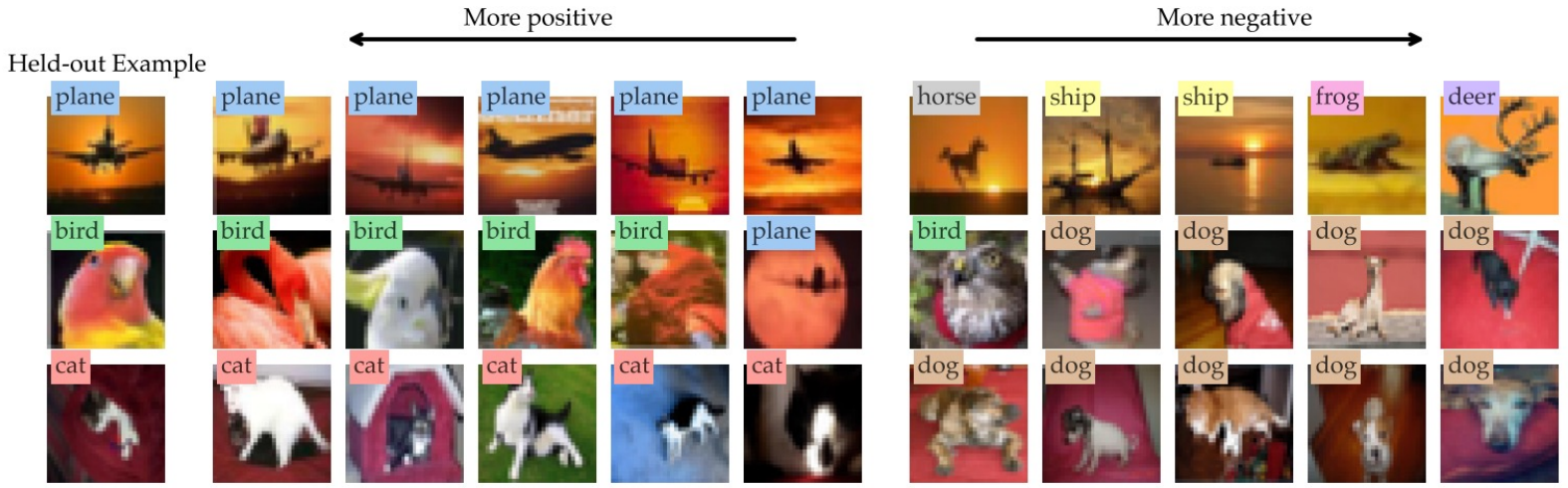


Pang Wei Koh, Percy Liang "Understanding Black Box Predictions via Influence Functions"

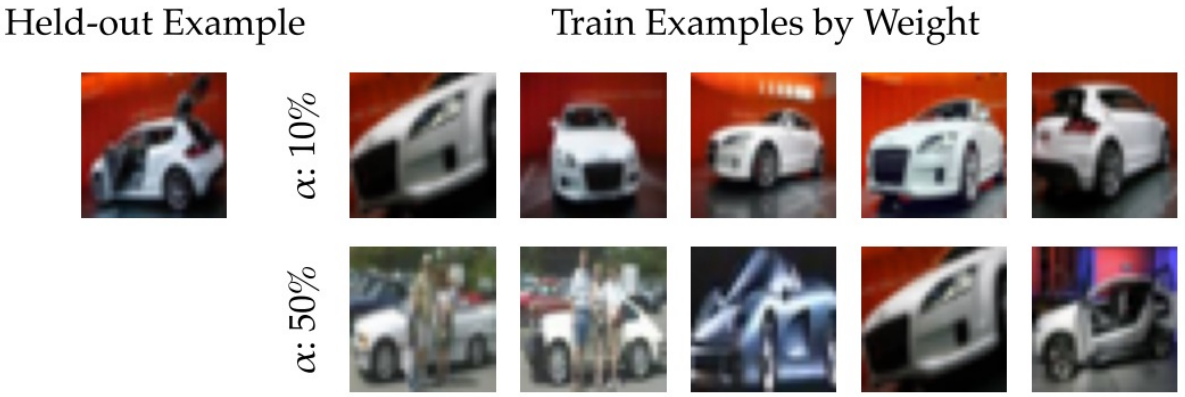
Linear datamodels are enough



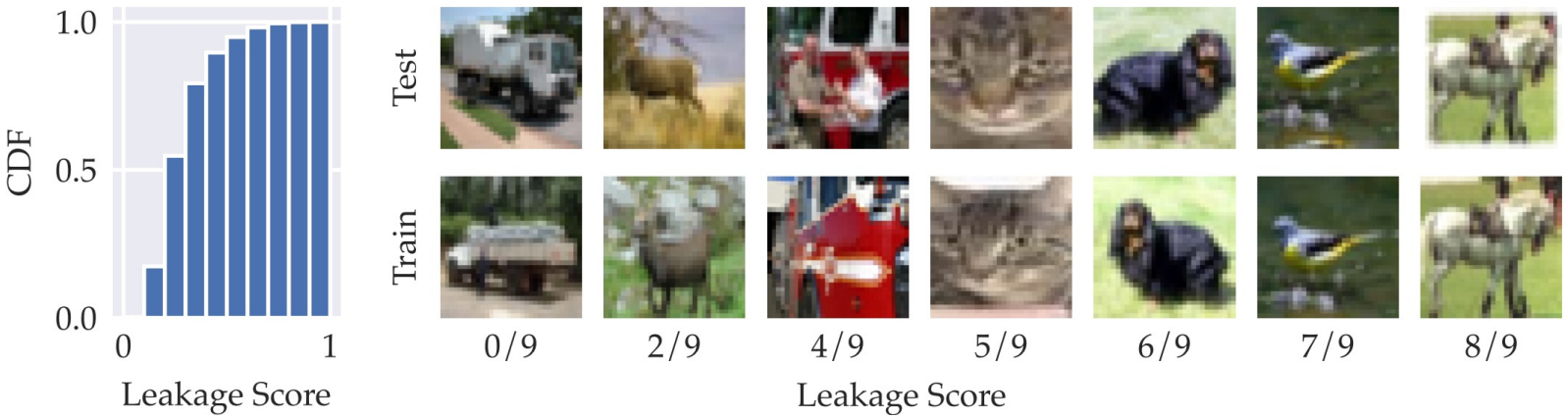
Datamodels show similar images



Effect of subset size

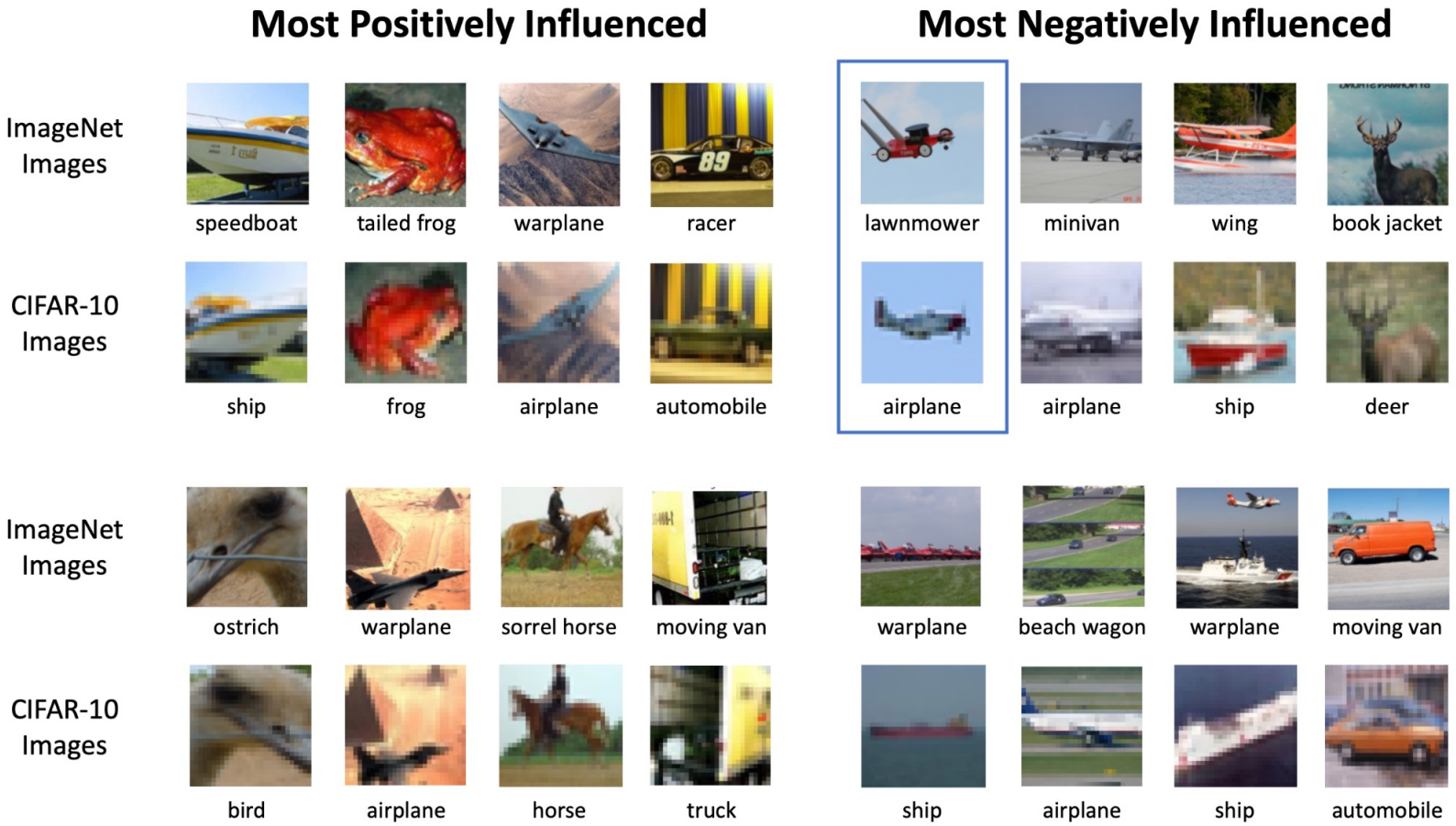


Data leakage



Clustering datamodel weights

Similar for transfer learning



Subpopulations in transfer

CIFAR10 datapoints with high influence from ImageNet Ostriches