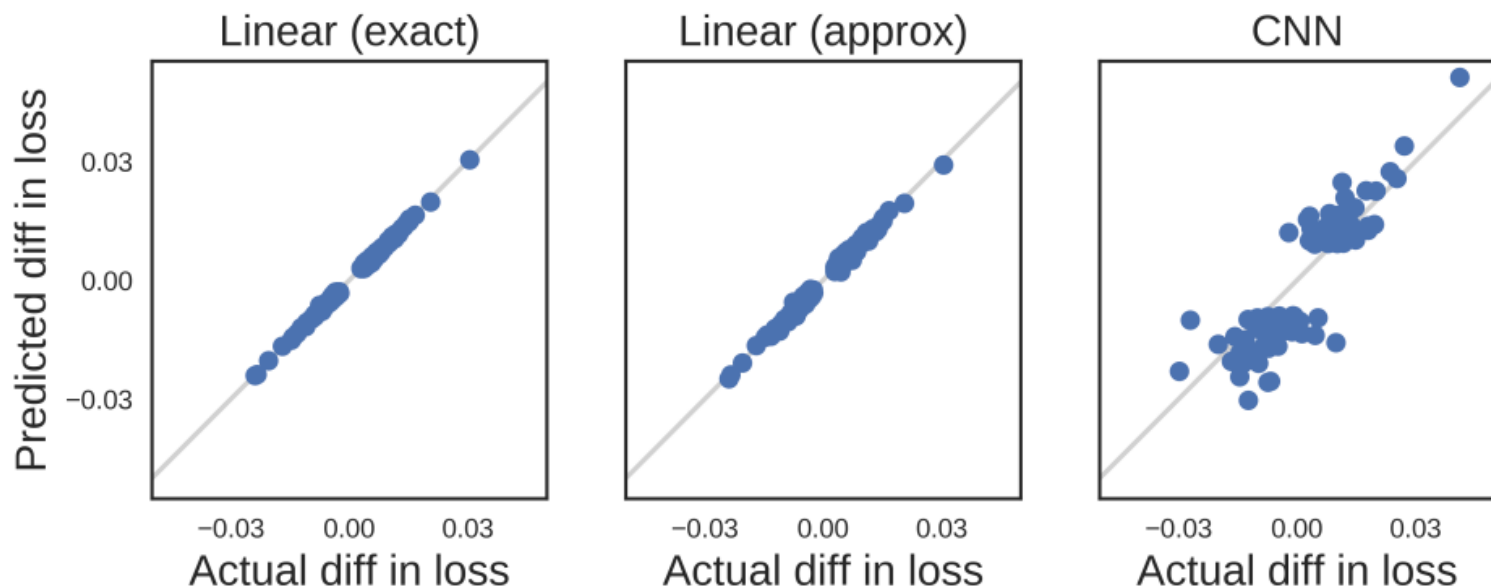


# Scientific discovery

Eric Wong

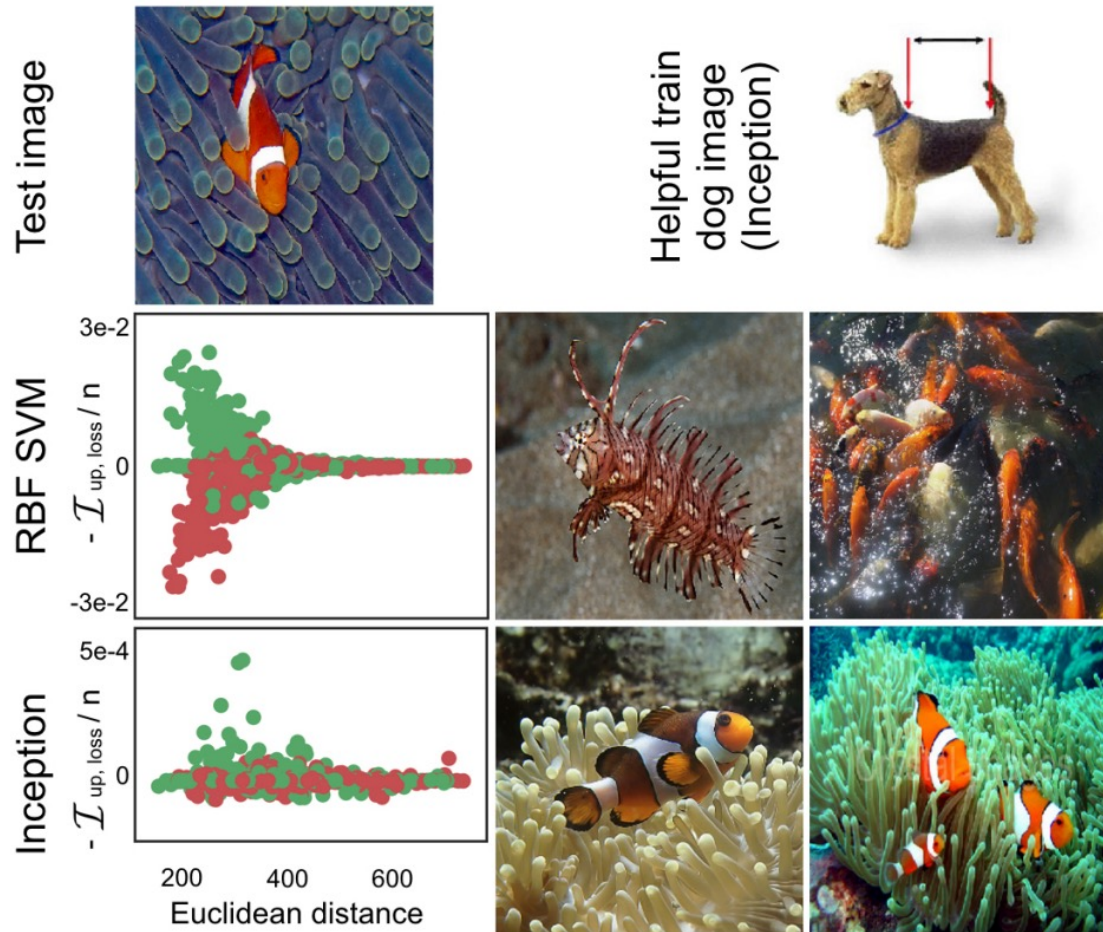
10/20/2022

# Influence functions approximate deletion



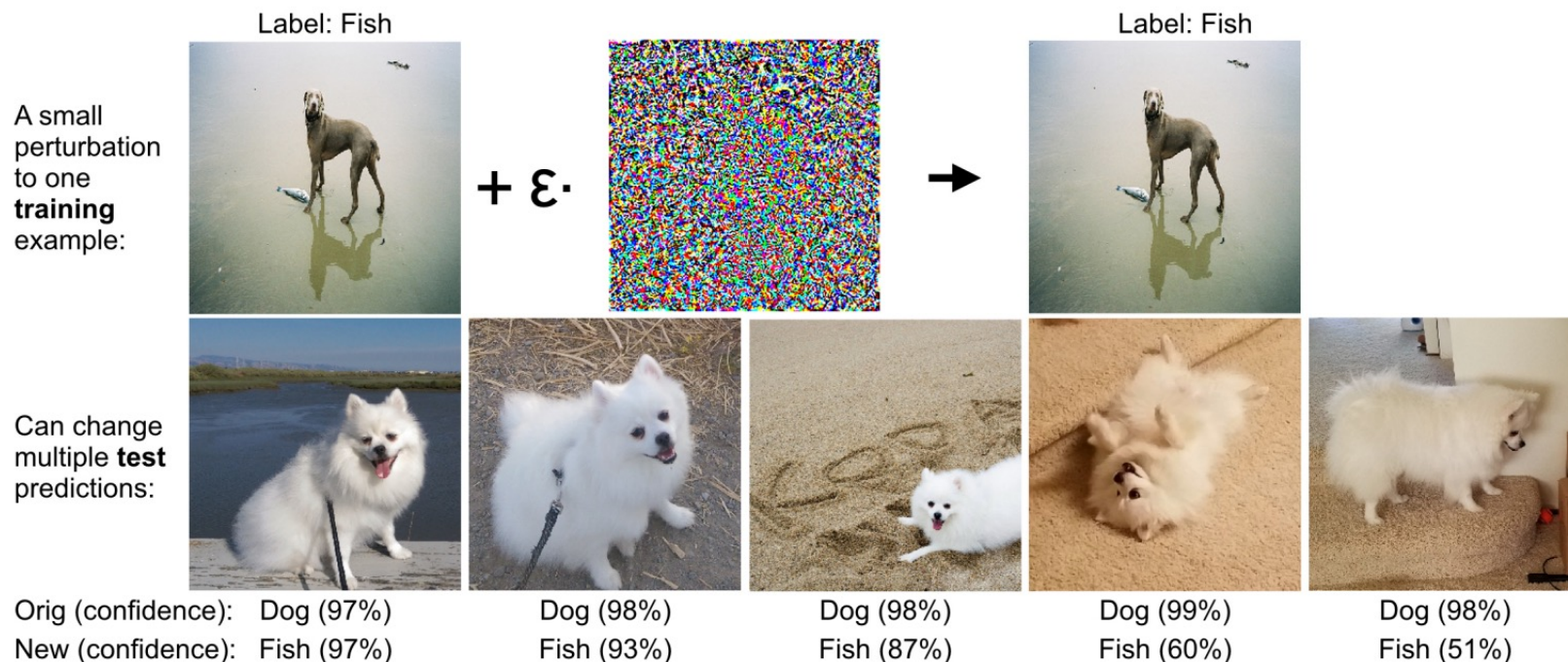
Pang Wei Koh, Percy Liang "Understanding Black Box Predictions via Influence Functions"

# Influential examples

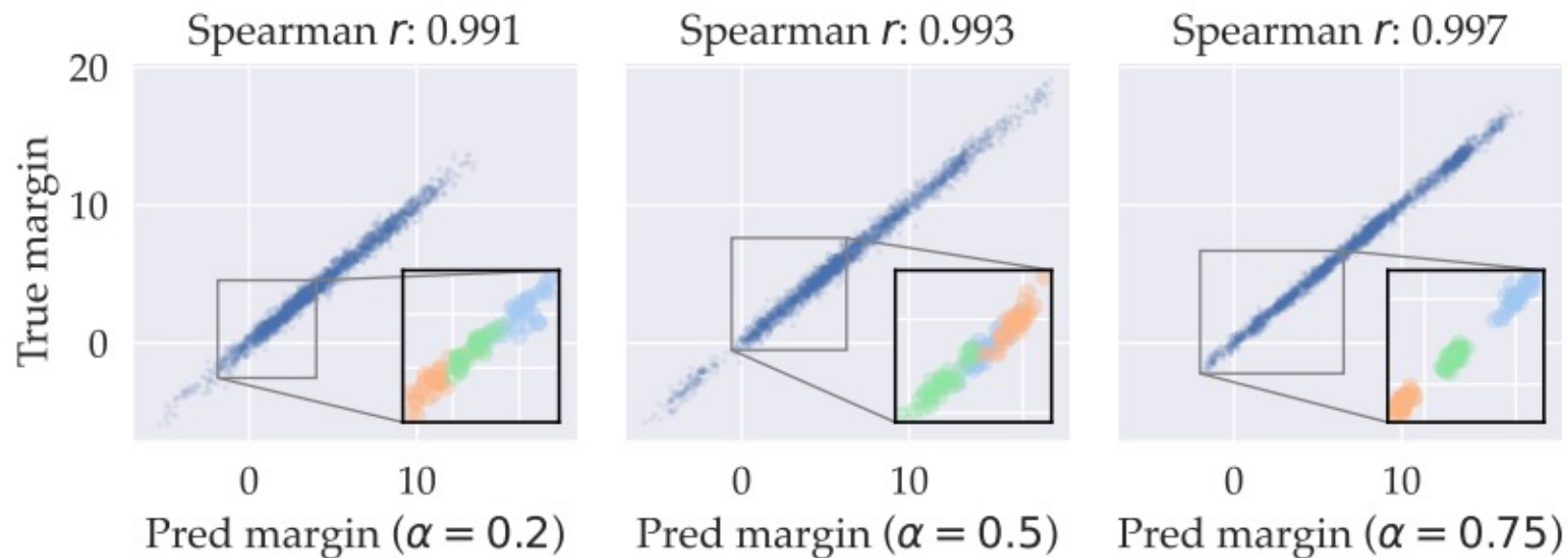


Pang Wei Koh, Percy Liang "Understanding Black Box Predictions via Influence Functions"

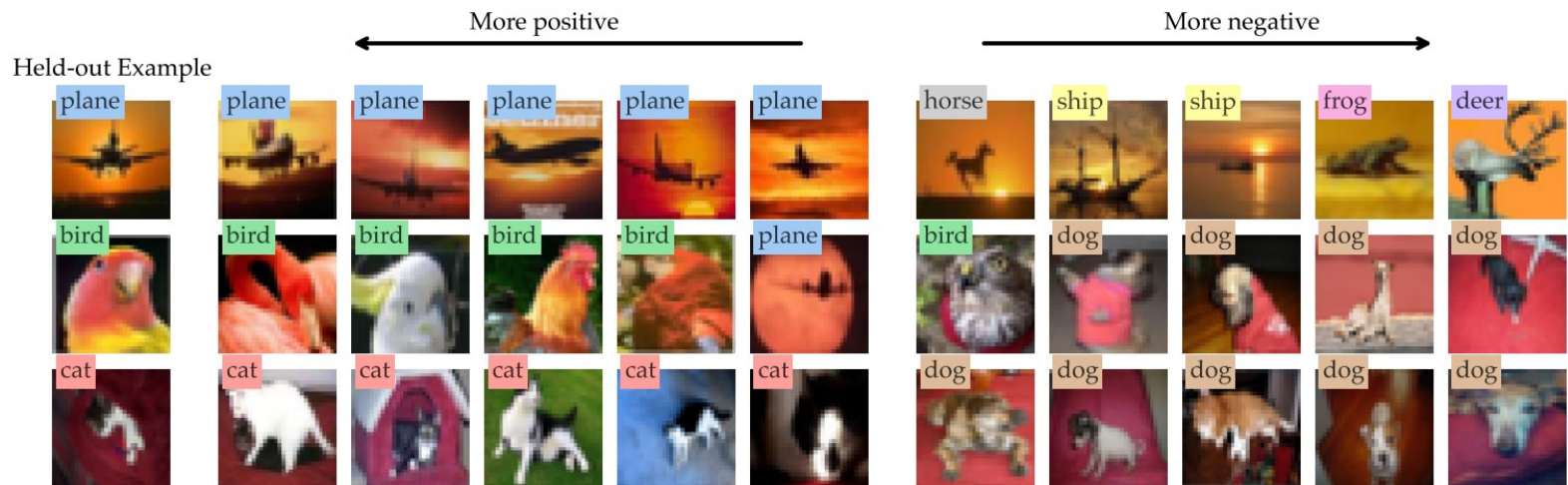
# Attack influential examples



# Linear datamodels are enough



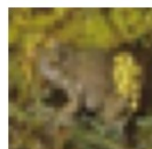
# Data models show similar images



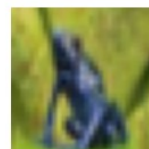
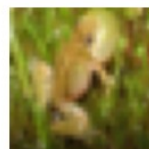
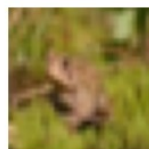


# Effect of subset size

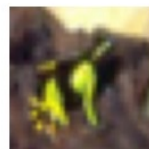
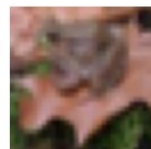
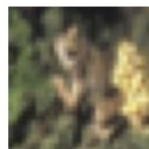
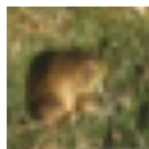
Held-out Example



$\alpha$ : 10%

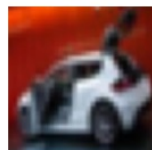


$\alpha$ : 50%

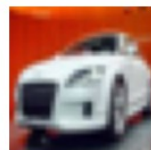
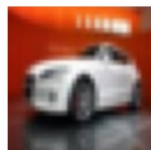
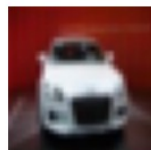


Train Examples by Weight

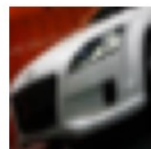
Held-out Example



$\alpha$ : 10%

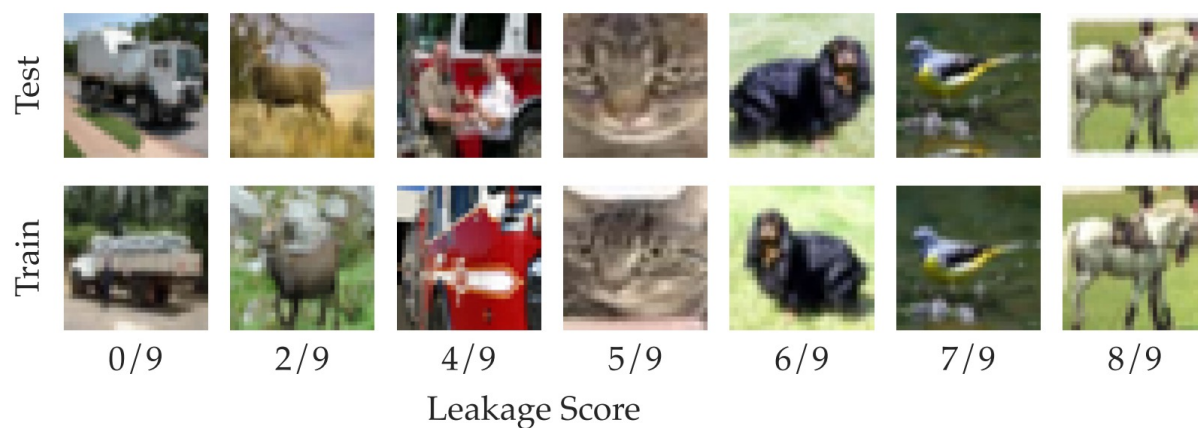
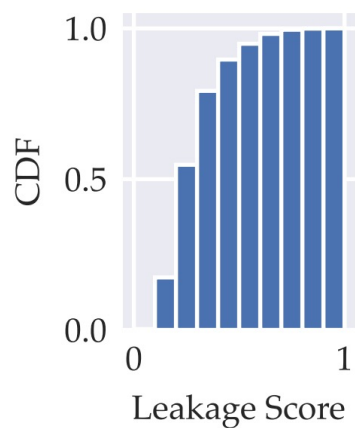


$\alpha$ : 50%



Train Examples by Weight

# Data leakage





# Clustering datamodel weights



# Similar for transfer learning

## Most Positively Influenced

ImageNet  
Images



speedboat



tailed frog



warplane

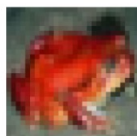


racer

CIFAR-10  
Images



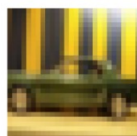
ship



frog

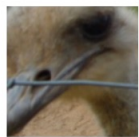


airplane



automobile

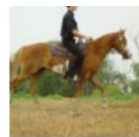
ImageNet  
Images



ostrich



warplane

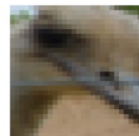


sorrel horse

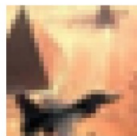


moving van

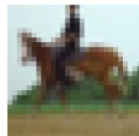
CIFAR-10  
Images



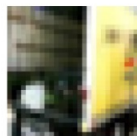
bird



airplane

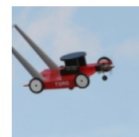


horse



truck

## Most Negatively Influenced



lawnmower



minivan



wing



book jacket



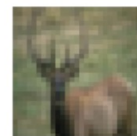
airplane



airplane



ship



deer



warplane



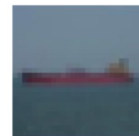
beach wagon



warplane



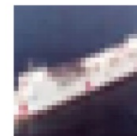
moving van



ship



airplane



ship



automobile

# Subpopulations in transfer

CIFAR10 datapoints with high influence from ImageNet Ostriches

