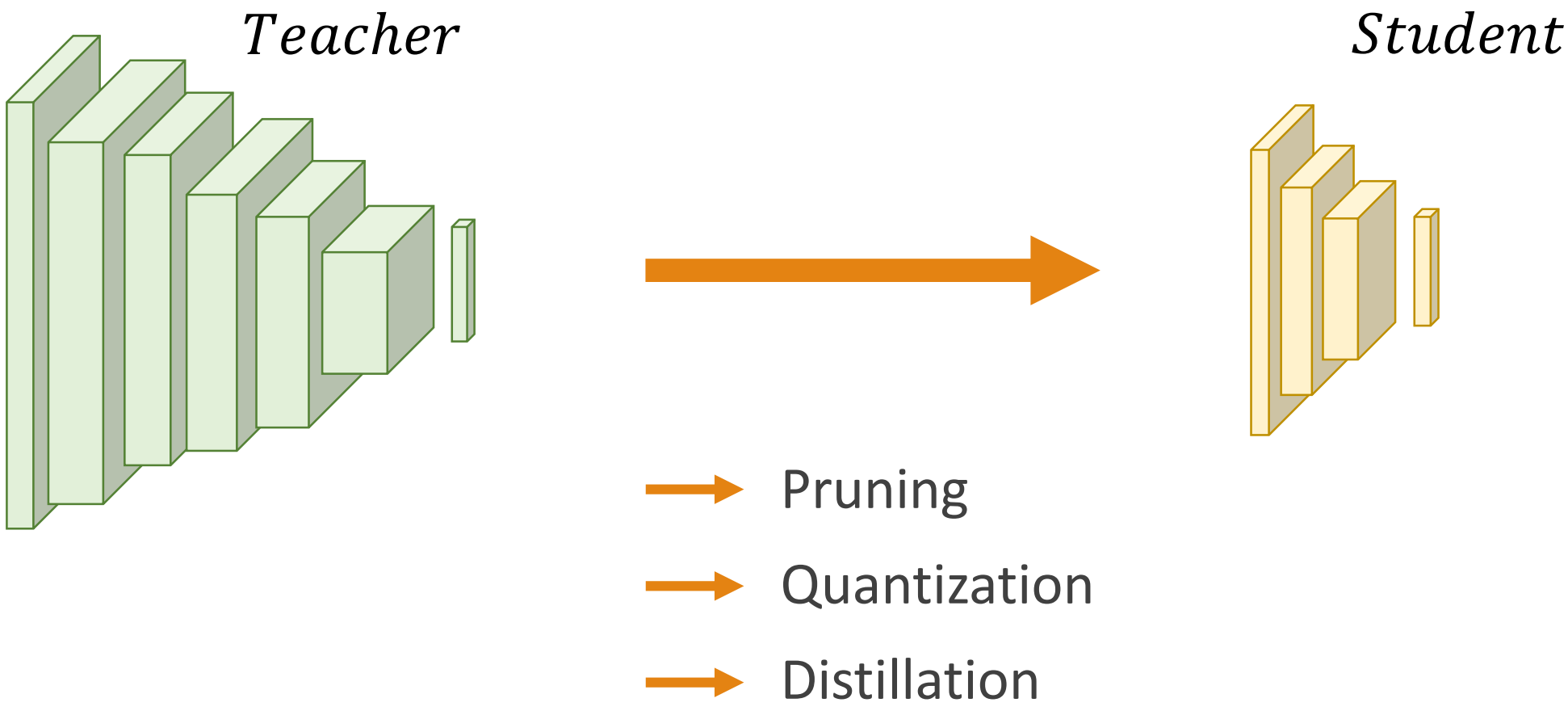
Zero-shot Knowledge Transfer

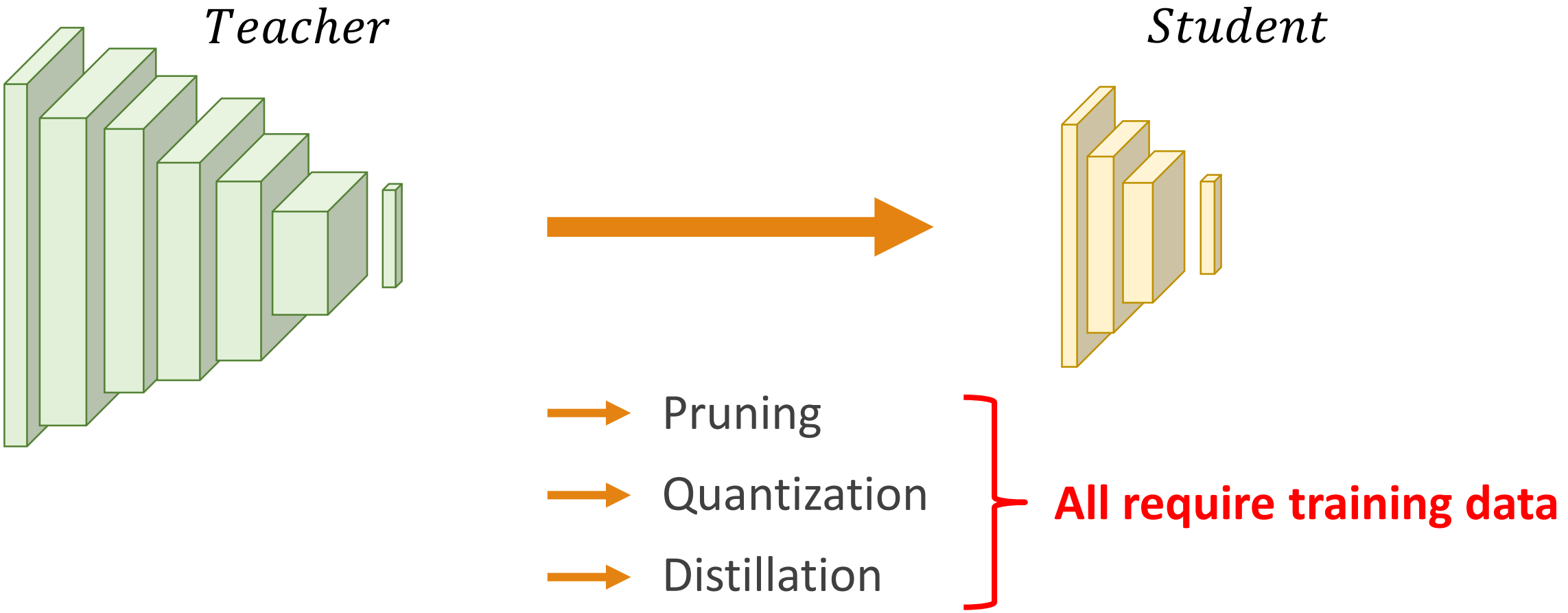
VIA ADVERSARIAL BELIEF MATCHING

Paul Micaelli and Amos Storkey

Model Compression



Model Compression

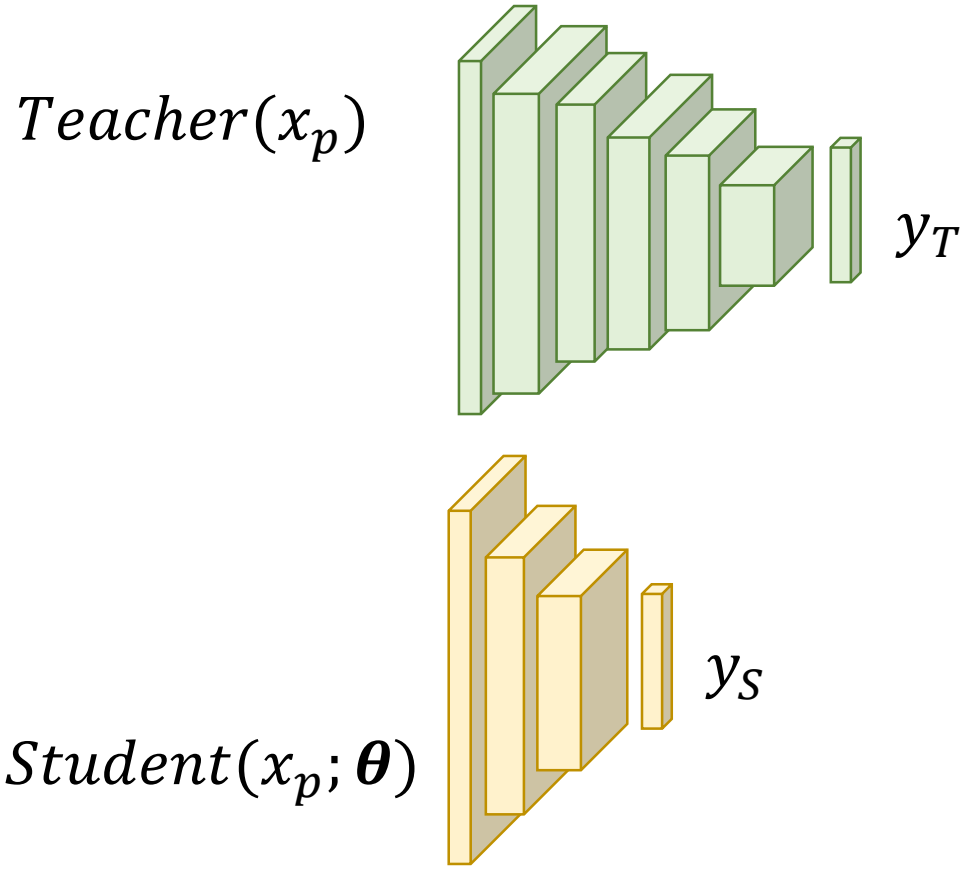


But data may not be available

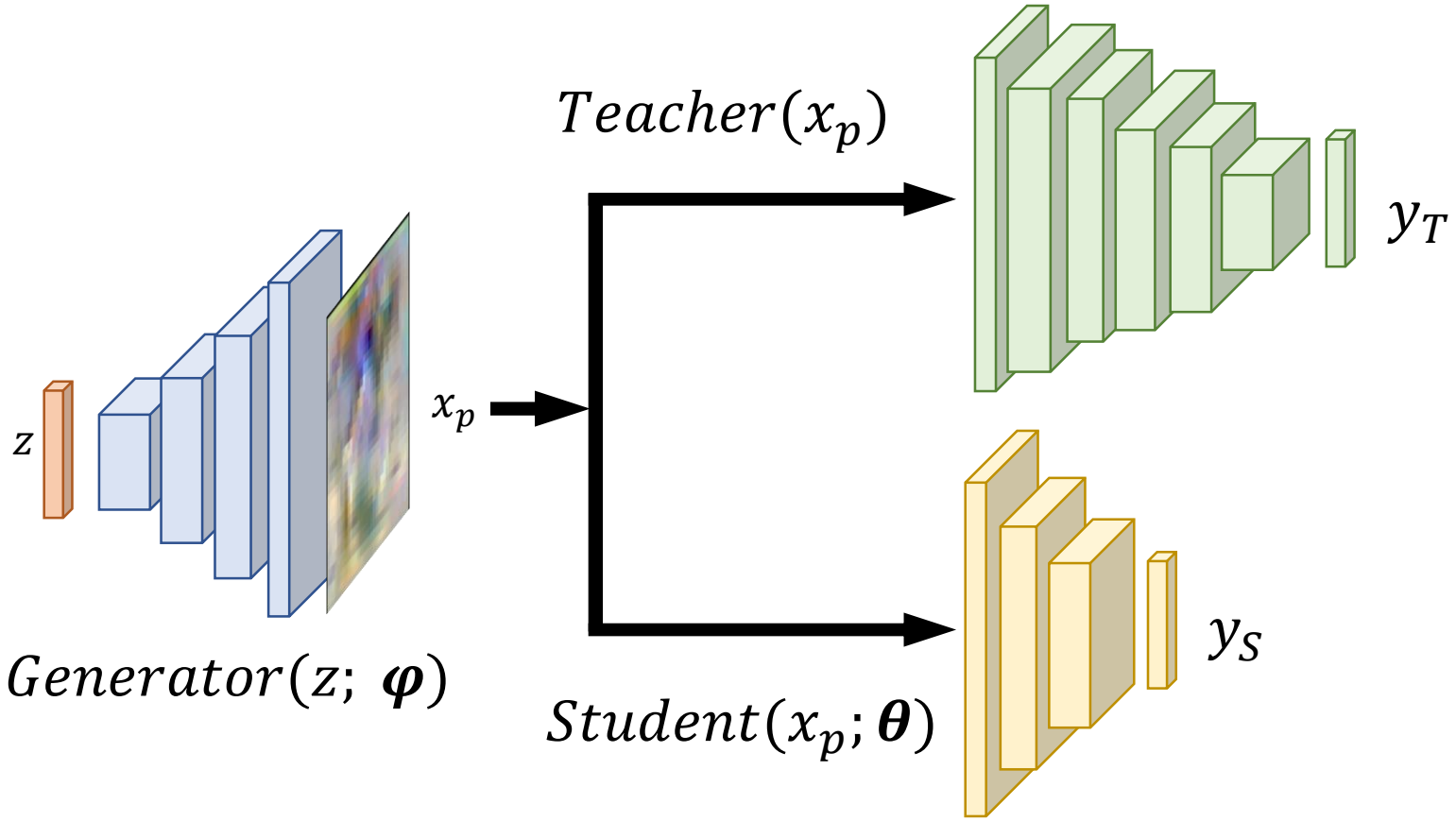
- Sometimes models are released without training data:
  - Privacy
  - Property
  - Size
  - Transience

- We need to do **model compression without a dataset**

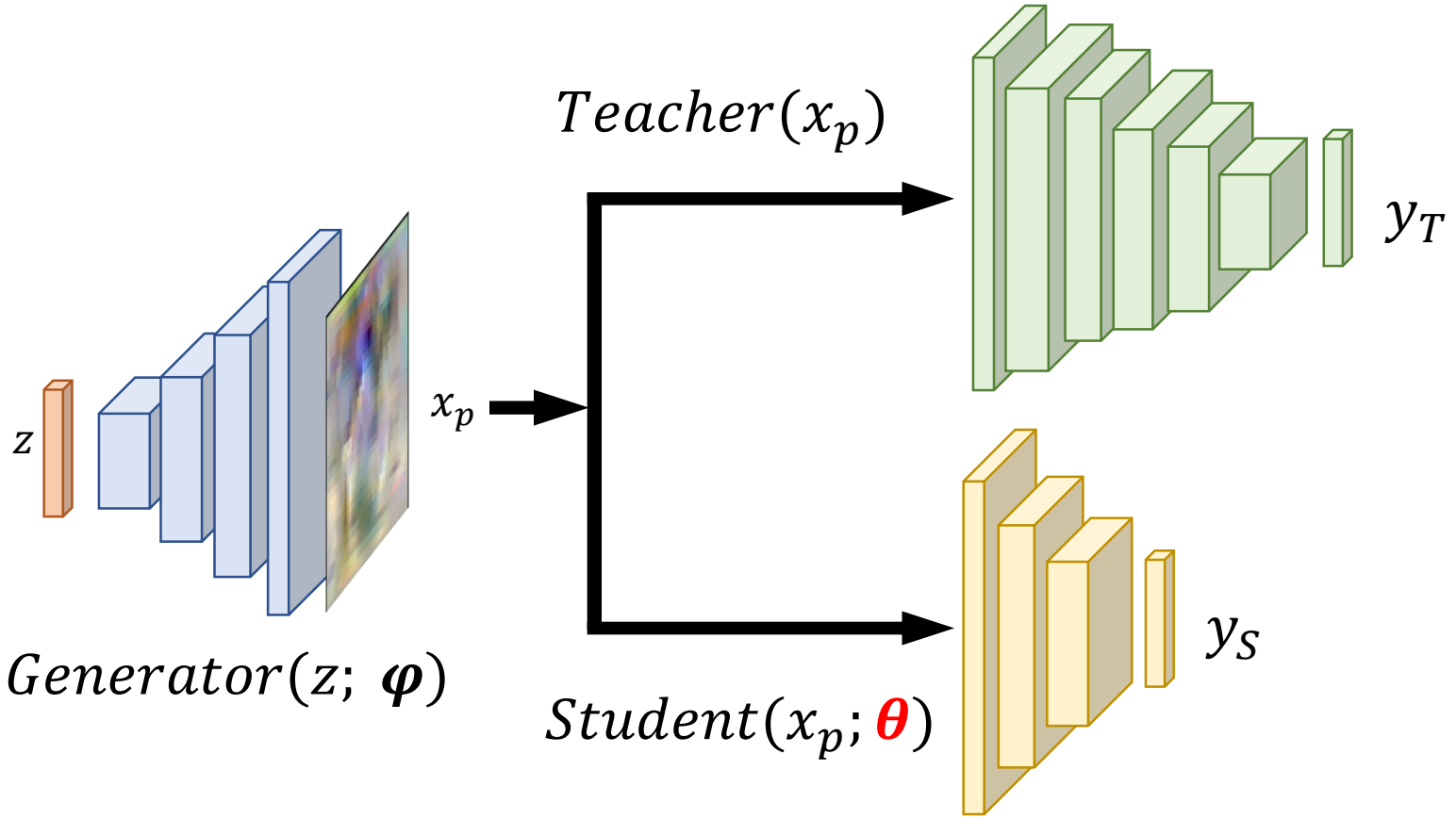
Zero-shot knowledge transfer



Zero-shot knowledge transfer

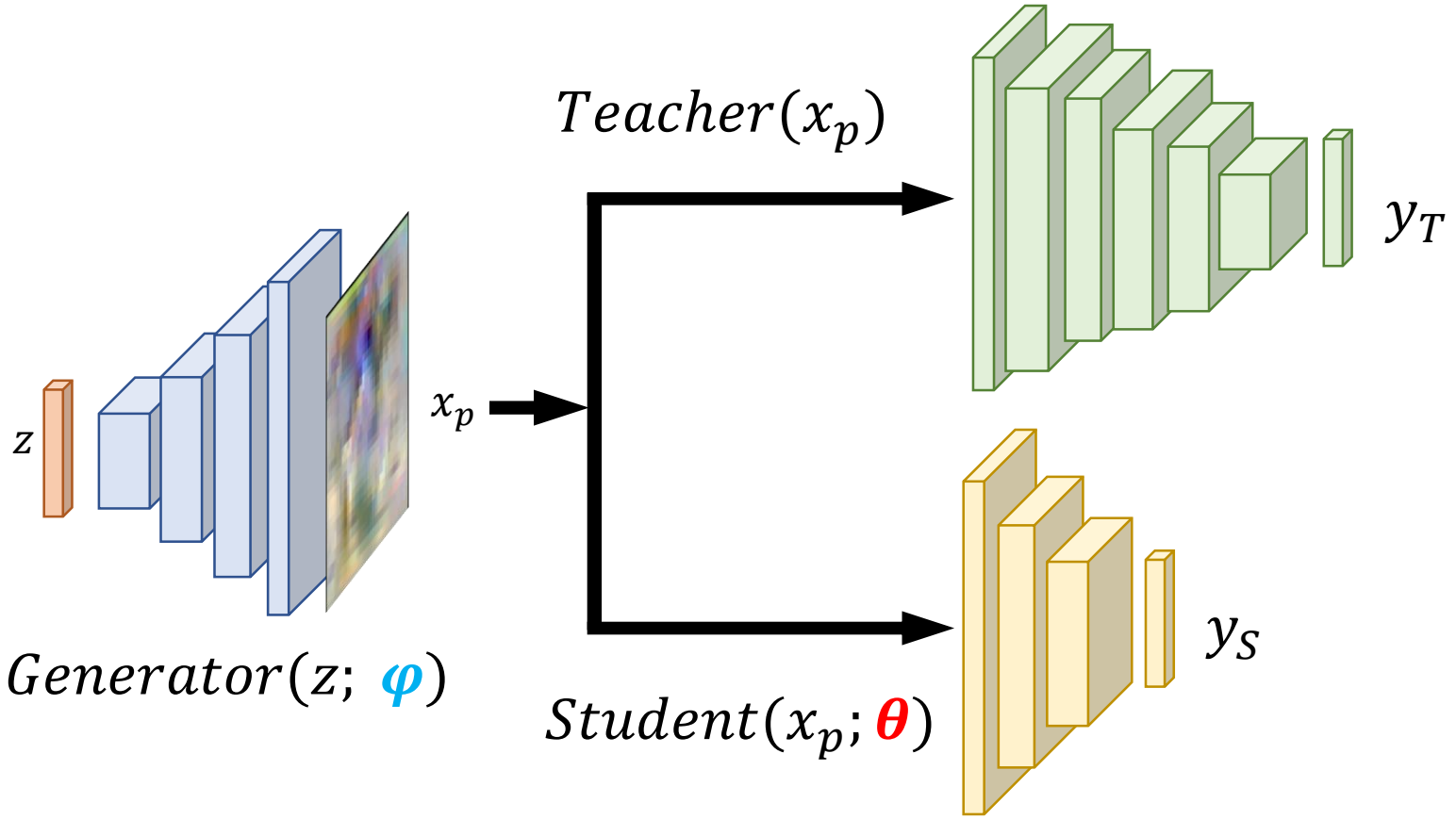


Zero-shot knowledge transfer



$$\min_{\boldsymbol{\theta}} \ D_{KL}(y_T, y_S)$$

Zero-shot knowledge transfer

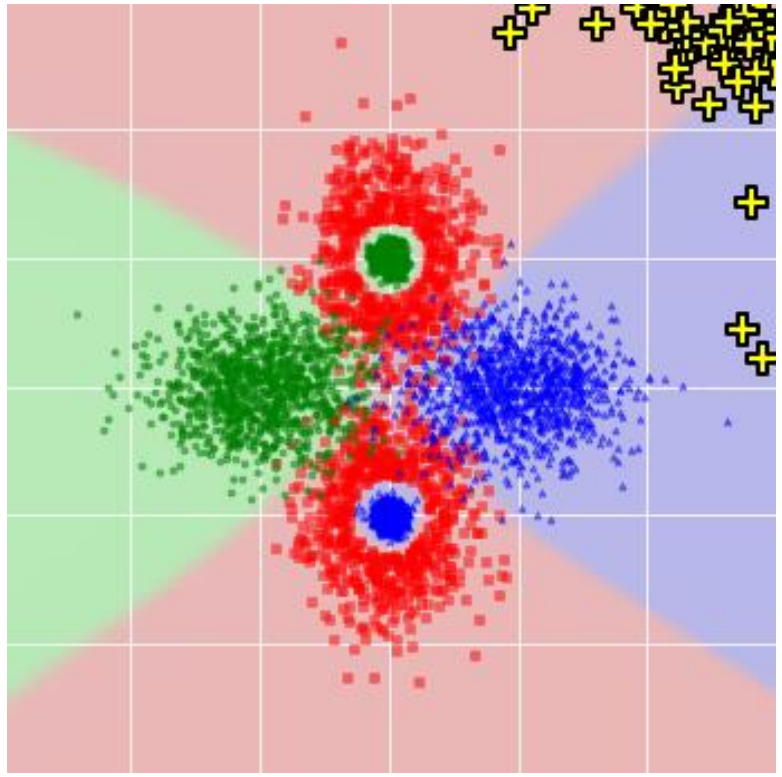


$$\max_{\boldsymbol{\varphi}} \min_{\boldsymbol{\theta}} D_{KL}(y_T, y_S)$$

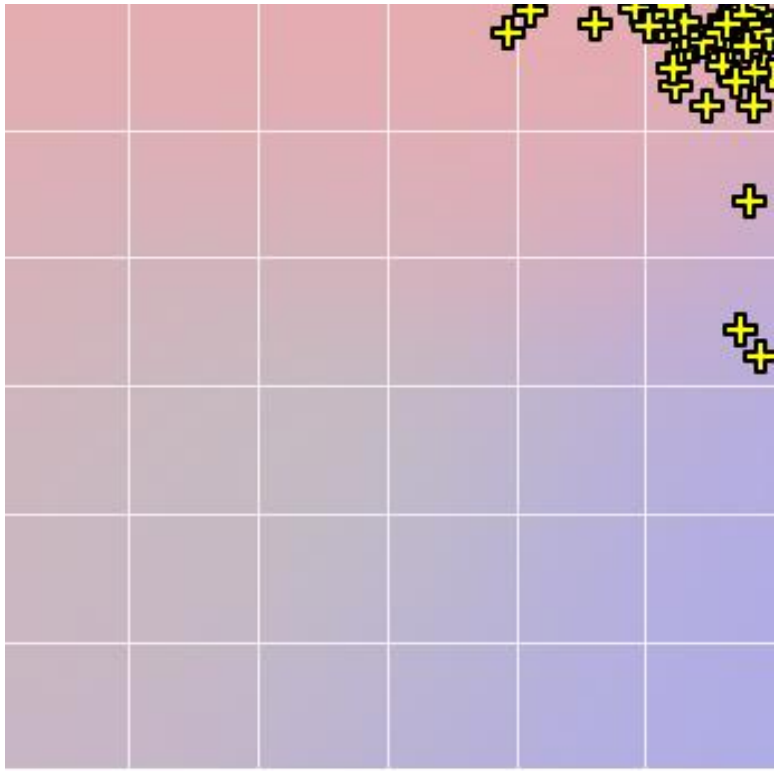
Toy problem visualization

$t = 0$

Teacher



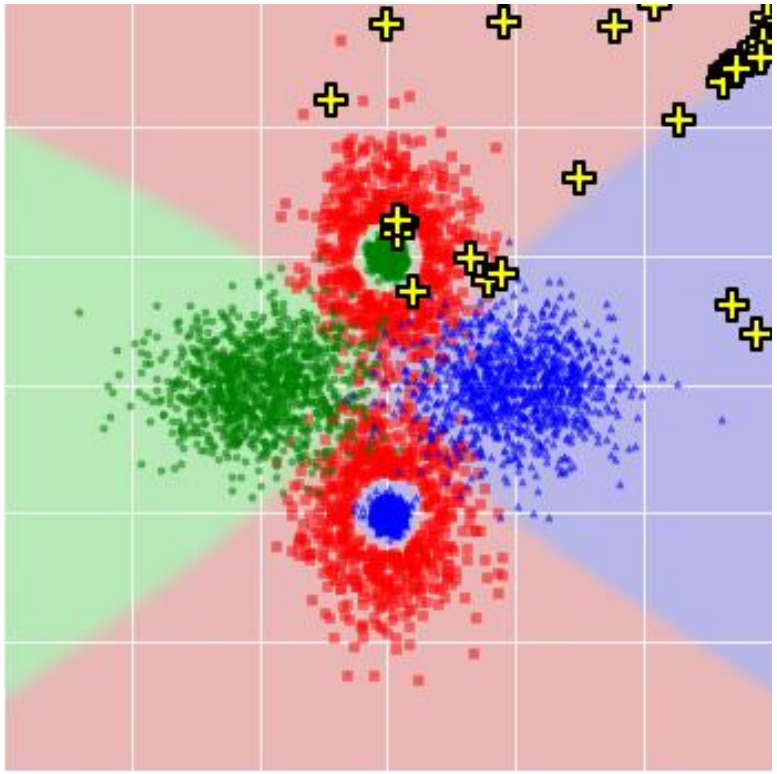
Student



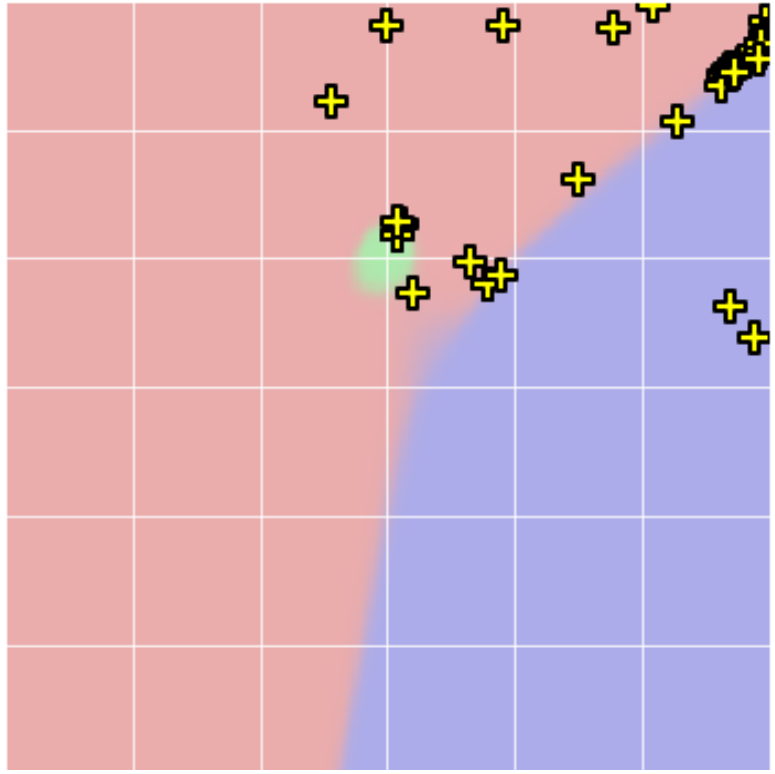
Toy problem visualization

$t = 1$

Teacher



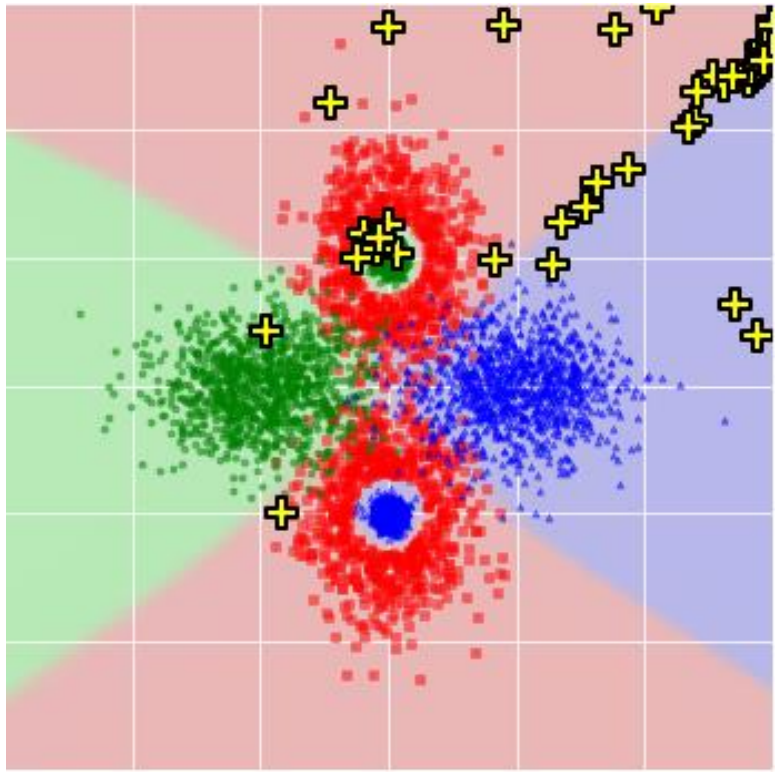
Student



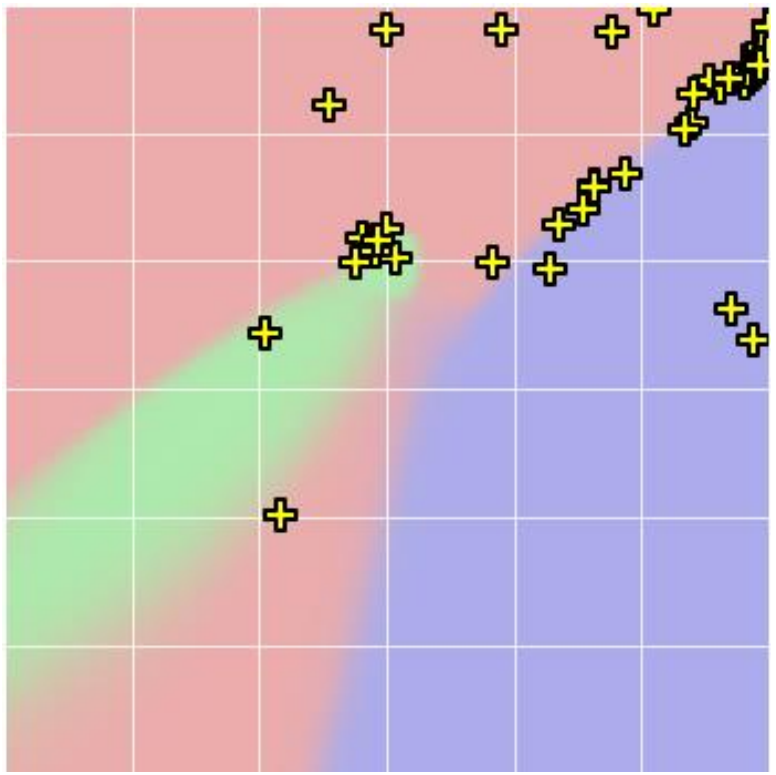
Toy problem visualization

$t = 2$

Teacher



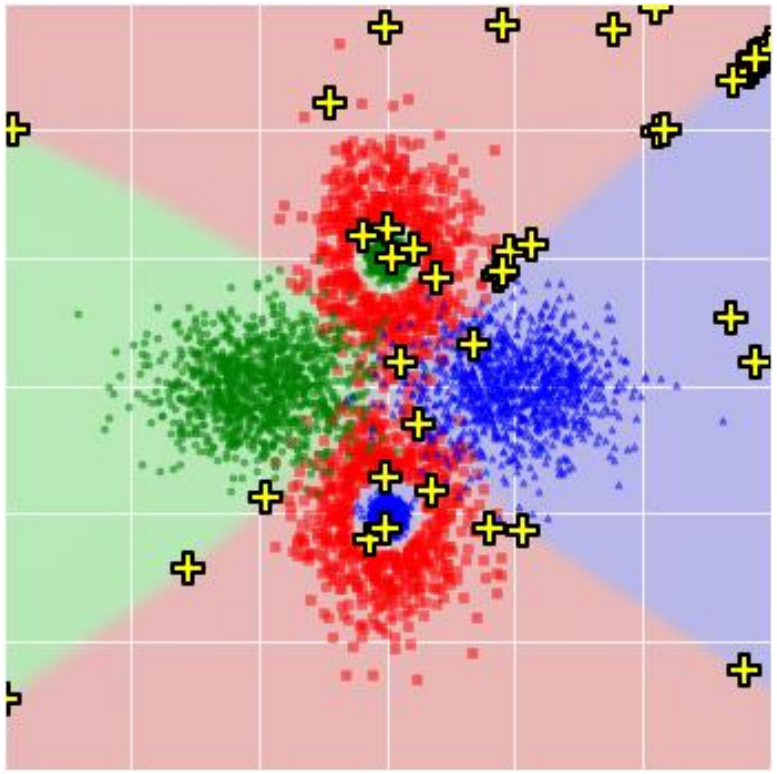
Student



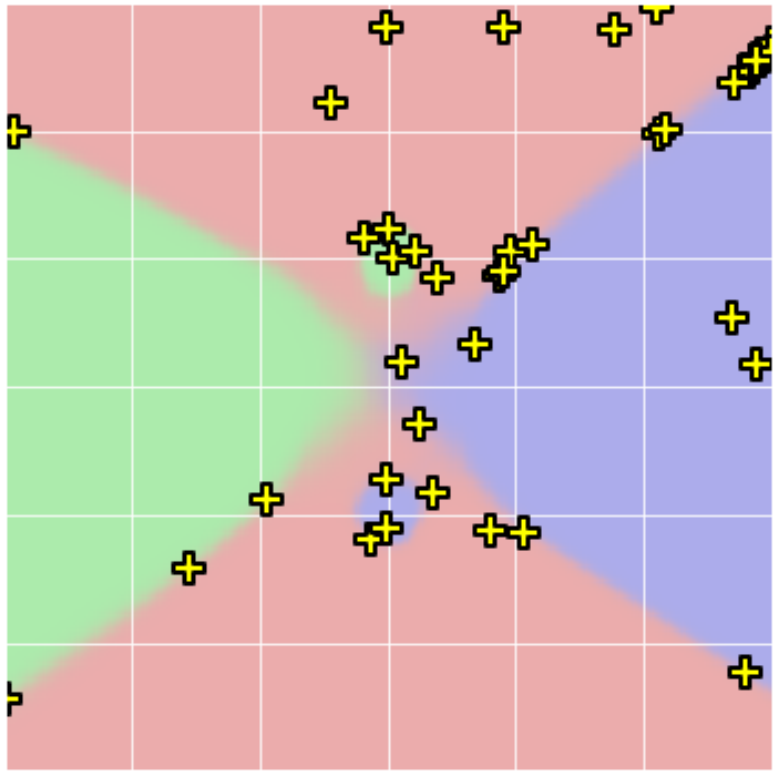
Toy problem visualization

$t = 3$

Teacher

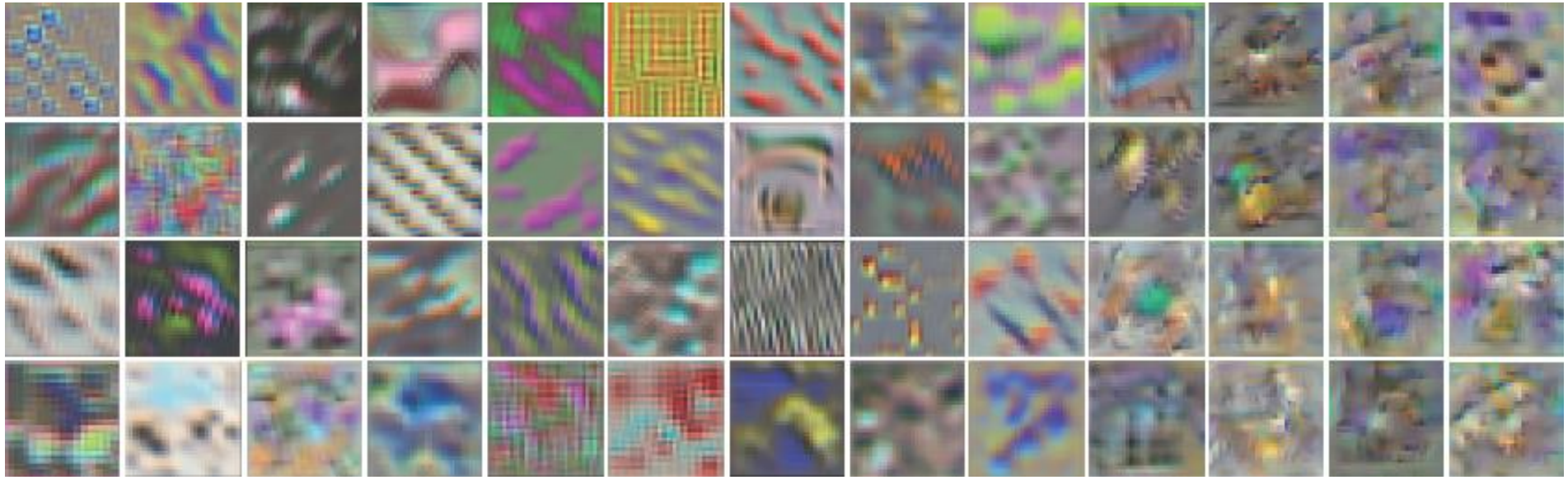


Student



CIFAR-10 pseudo images

- $\approx 90\%$ test accuracy for a WRN-16-2
- Pseudo images don't look like real images, but matching the teacher on them leads to matching it more generally



Mean Transition Error (MTE)

- We propose a metric to measure decision boundary match between two networks near real data
- MTE is higher for our student than for the baseline student even though ours has never matched the teacher on real data

	Zero-shot (Ours)	KD+AT
SVHN	0.09	0.64
CIFAR-10	0.22	0.68

Summary

- We can compress a teacher to a student without using a dataset
- This is done by adversarially finding regions of input space where the student and teacher disagree