[1] BOSTON UNIVERSITY

[2] Carnegie Mellon University

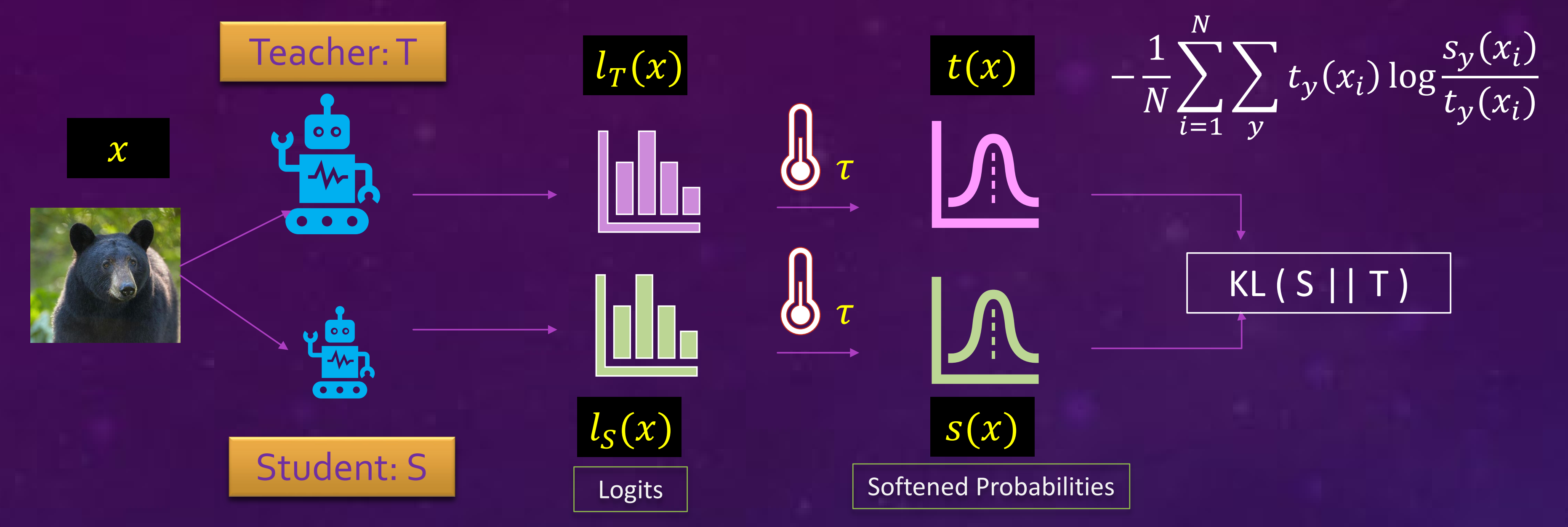
SCAFFOLDING A STUDENT TO INSTILL KNOWLEDGE

ANIL KAG [1], DURMUS ALP EMRE ACAR [1], ADITYA GANGRADE [2], VENKATESH SALIGRAMA [1]

ICLR 2023

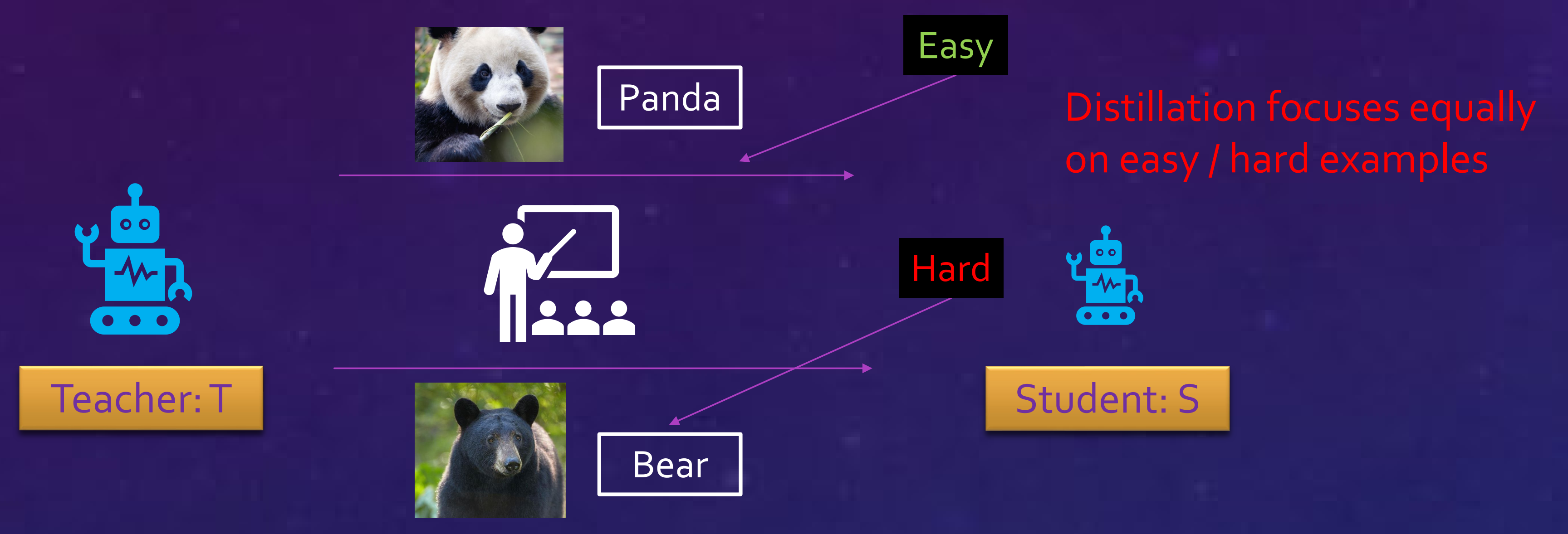
Problem: Teaching a Tiny Model

- Knowledge Distillation Setup

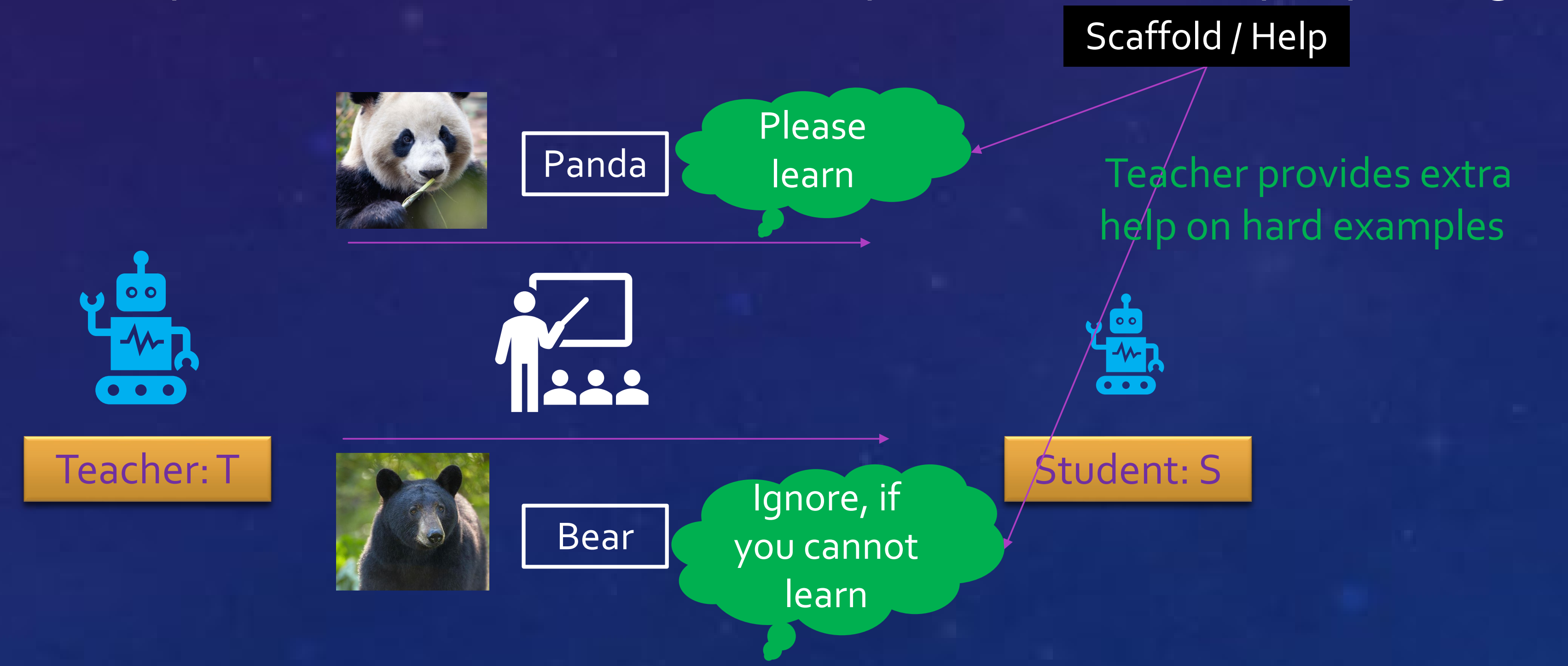


$$-\frac{1}{N}\sum_{i=1}^{N}\sum_{y} t_y(x_i)\log\frac{s_y(x_i)}{t_y(x_i)}$$

- Tiny Models have limited capacity

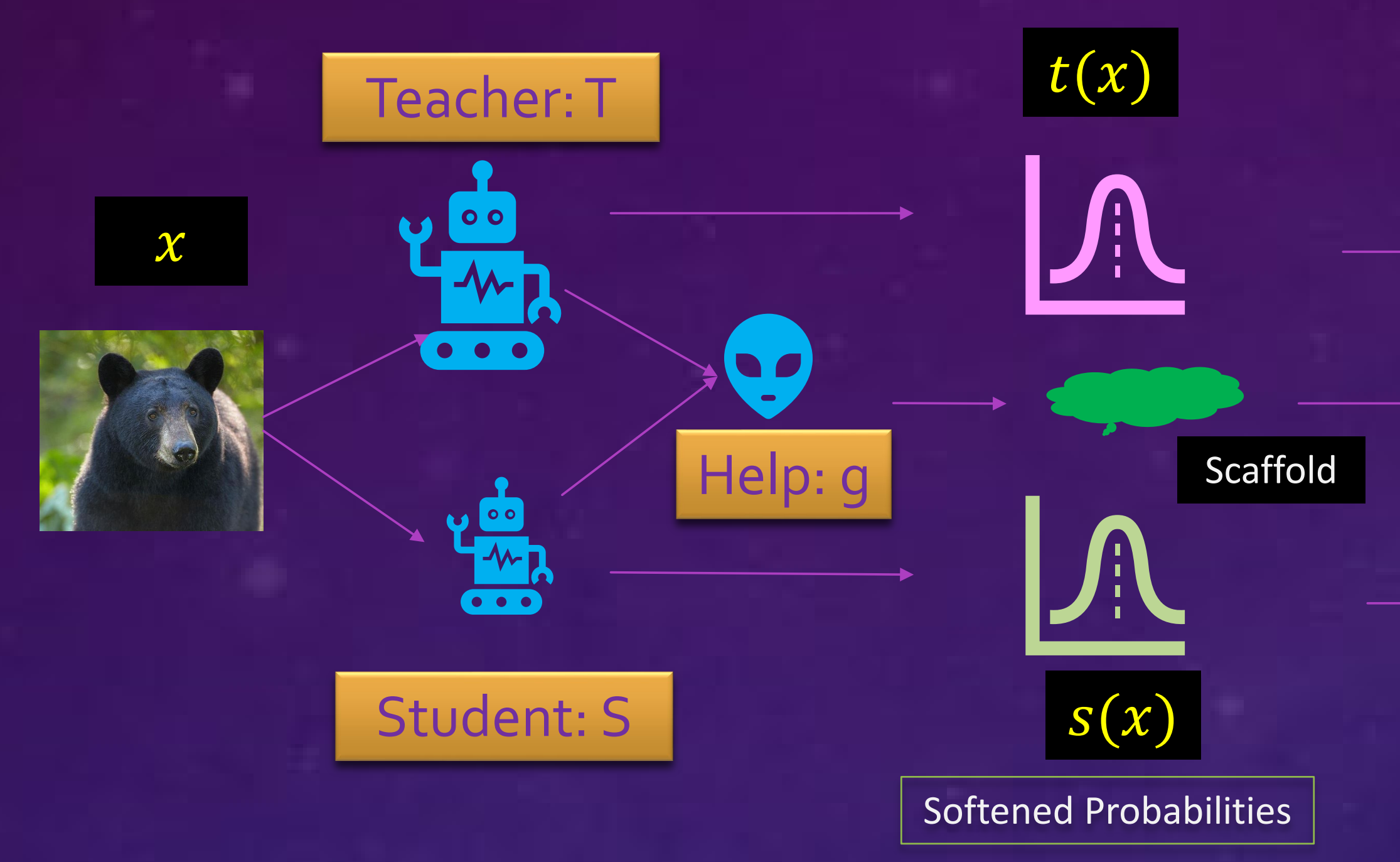


- Our Proposal : Scaffold hard-to-learn inputs for student by exploiting teacher


Scaffold / Help
Panda
Please learn
Teacher provides extra help on hard examples
Teacher: T
Student: S
Bear
Ignore, if you cannot learn

DiSK: Distilling Scaffolded Knowledge

- DiSK Setup



- ❑ *g is computationally cheap*
- ❑ *g is used only in training*

$$-\frac{1}{N}\sum_{i=1}^{N}\sum_{y} t_y(x_i)\log\frac{s_y(x_i)+g(x_i)}{t_y(x_i)}$$

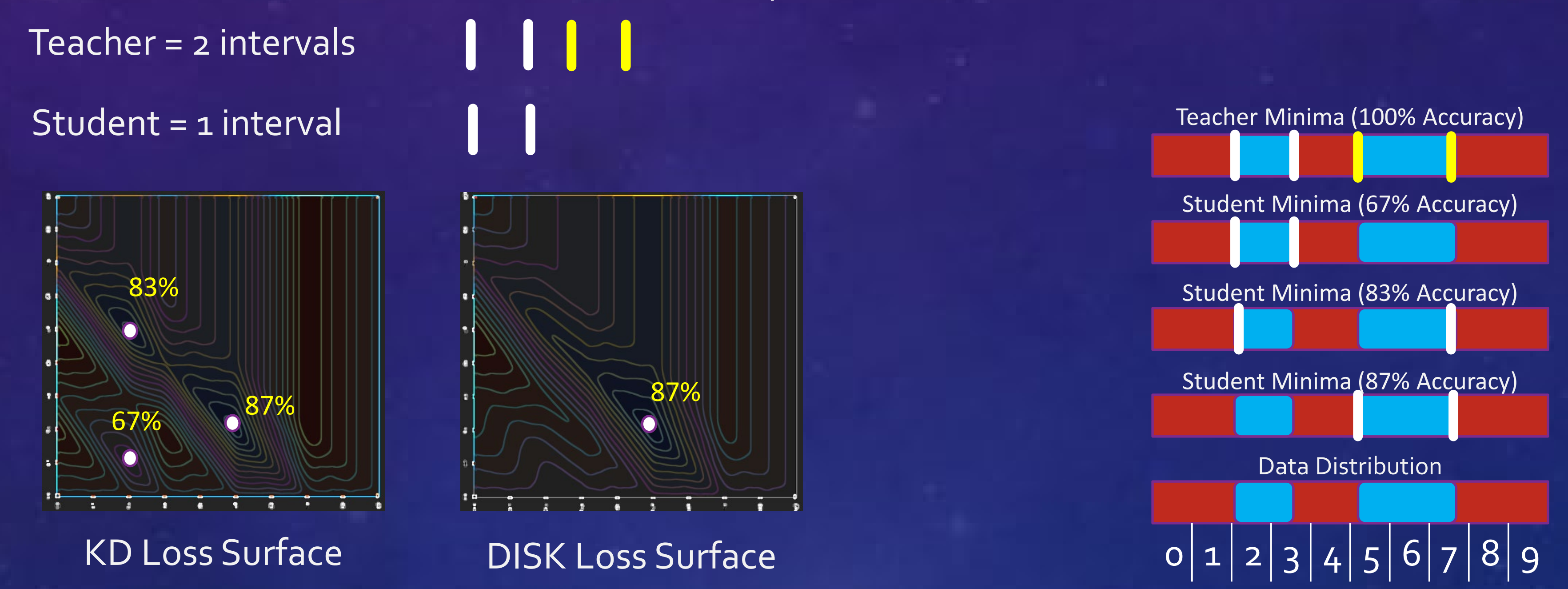
KL (S + g || T) + Budget(g)

Avoid trivial solution
(help on all examples)

$$\max(0,\ -\frac{1}{N}\sum_{i=1}^{N} g(x_i)\log s_{y_i}(x_i)\ -\delta)$$

- DiSK smoothens the student's loss-landscape, often eliminating suboptimal minima

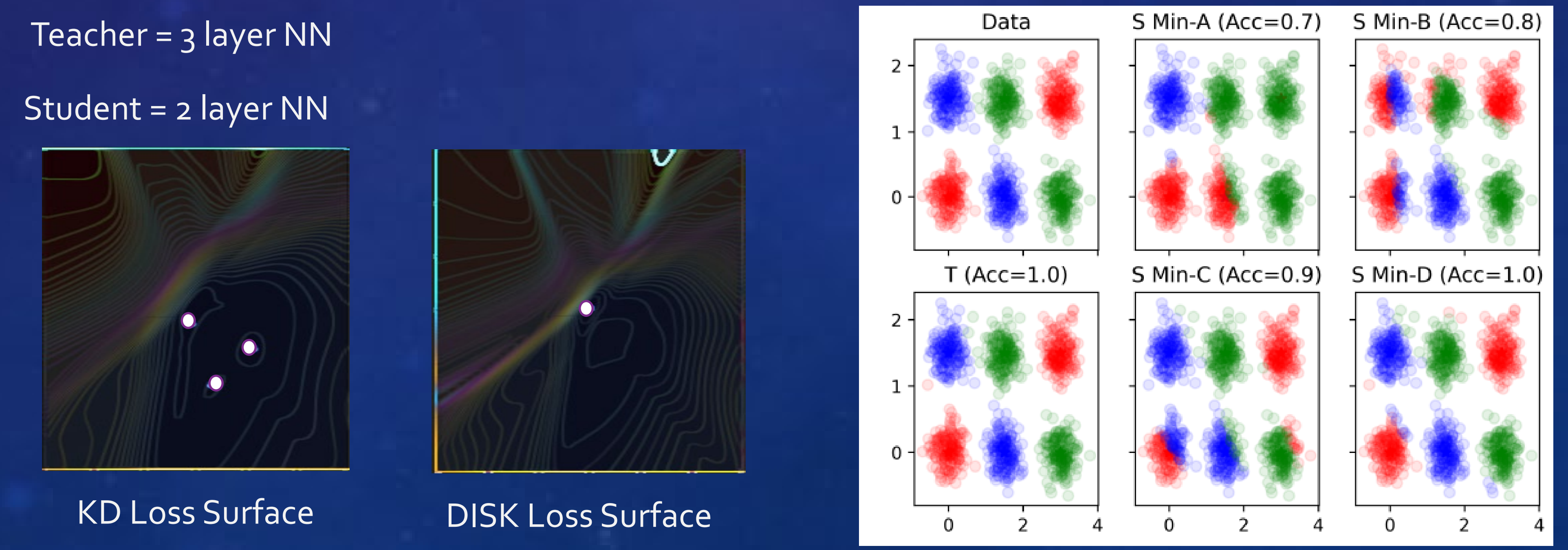
Example: 1D Intervals



KD Loss Surface

DISK Loss Surface

Example: 2D Gaussians



KD Loss Surface

DISK Loss Surface

Empirical Evaluation

- CIFAR-100 (Tiny Students): Up to 4% higher accuracy compared to KD

Teacher	Teacher MACs	Student	Student MACs	CE	KD	DiSK
ResNet10-l	64M	Resnet10-s	4M	52.16	54.92	**58.14**
71.99%		Resnet10-m	16M	65.24	66.96	**70.03**
ResNet18	555M	Resnet10-s	4M	52.16	55.76	**58.11**
76.56		Resnet10-m	16M	65.24	68.09	**69.86**

- CIFAR-100 (Standard Students): Up to 2.5% higher accuracy compared to KD

Teacher	Teacher MACs	Student	Student MACs	CE	KD	DiSK
ResNet32x4	1083M	ShuffleNetV2	45M	73.74	79.13	**80.23**
81.45%		MobileNetV2x2	22M	69.24	76.05	**77.24**
Wide-ResNet	327M	ShuffleNetV2	45M	73.74	75.81	**78.33**
78.41		MobileNetV2x2	22M	69.24	73.92	**76.32**

- ImageNet-1K: More than 1% higher accuracy compared to KD

Teacher	Teacher MACs	Student	Student MACs	CE	KD	DiSK
ResNet50	4.12B	ResNet18	1.82B	69.73	71.29	**72.35**
ViT-Large	59.65B	ViT-Tiny	1.07B	75.45	76.61	**77.86**
ViT-Large	59.65B	DeiT-Tiny	1.07B	72.2	74.5	**75.59**

- DiSK can be integrated with other procedures such as feature matching

Teacher	Teacher MACs	Student	Student MACs	FitNet	SimKD	SimKD + DiSK
Wide-ResNet	327M	ResNet8x4	177M	75.02	76.75	**77.13**
78.41		Wide-ResNet-40-1	83M	74.17	75.56	**76.21**

- https://github.com/anilkagak2/DiSK_Distilling_Scaffolded_Knowledge