The Big Problem with Meta-Learning and How Bayesians Can Fix It

Chelsea Finn

Stanford

training data

Braque

Cezanne

test datapoint

By Braque or Cezanne?

How did you accomplish this?

Through previous experience.

How might you get a machine to accomplish this task?

Modeling image formation

Geometry

SIFT features, HOG features + SVM

Fine-tuning from ImageNet features
Domain adaptation from other painters

???

Fewer human priors,
more data-driven priors

Greater success.

Can we explicitly learn priors from previous experience
that lead to efficient downstream learning?

Can we learn to learn?

Outline

1. Brief overview of meta-learning
2. The problem: peculiar, lesser-known, yet ubiquitous
3. Steps towards a solution

How does meta-learning work? An example.

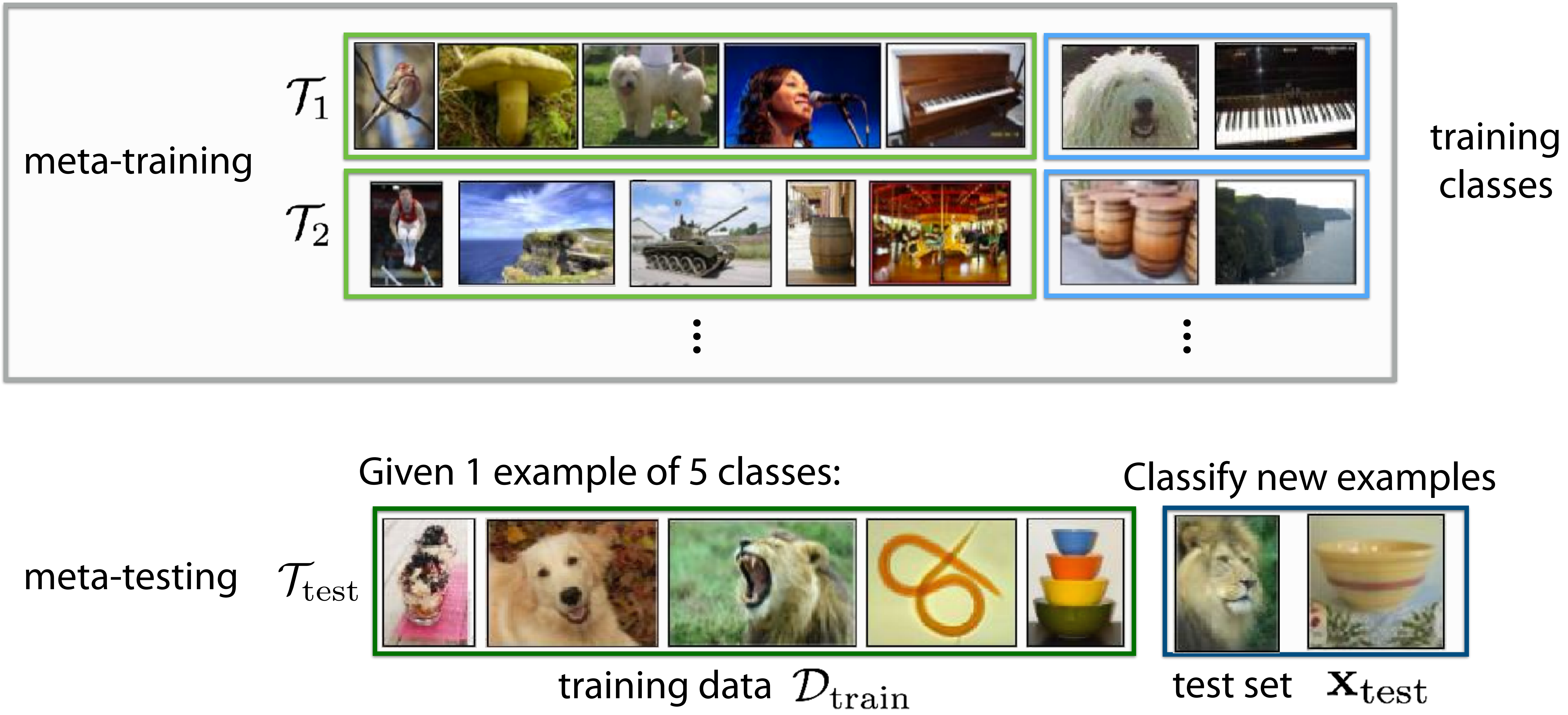
Given 1 example of 5 classes:

training data $\mathcal{D}_{\text{train}}$

Classify new examples

test set $\mathbf{x}_{\text{test}}$

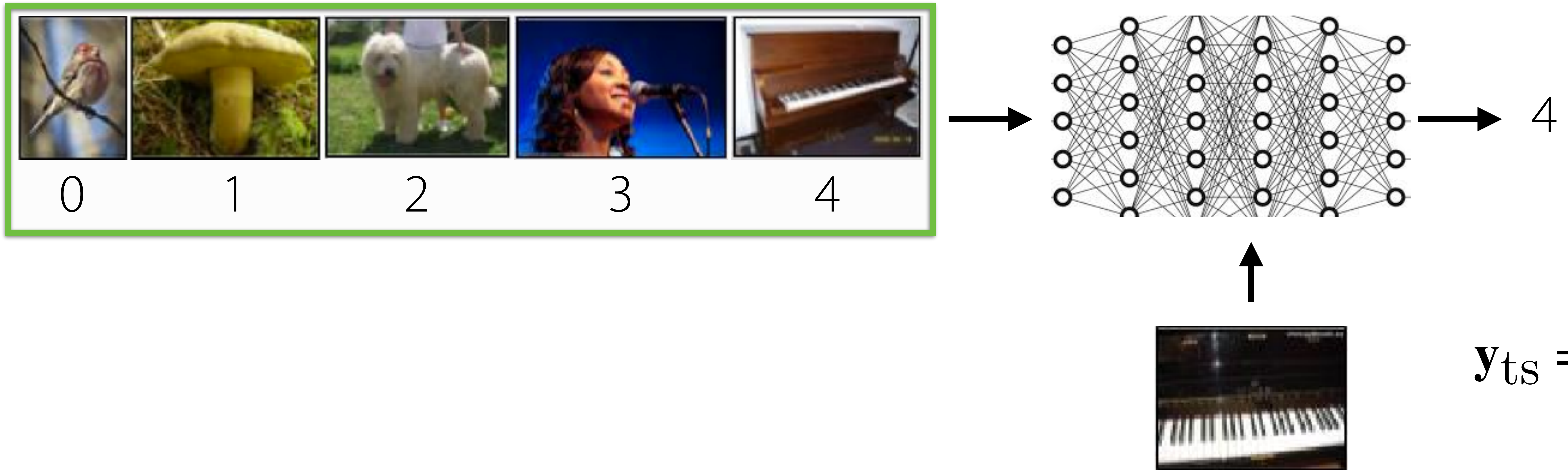
How does meta-learning work? An example.



How does meta-learning work?

$\mathcal{T}_1$

One approach: parameterize learner by neural network



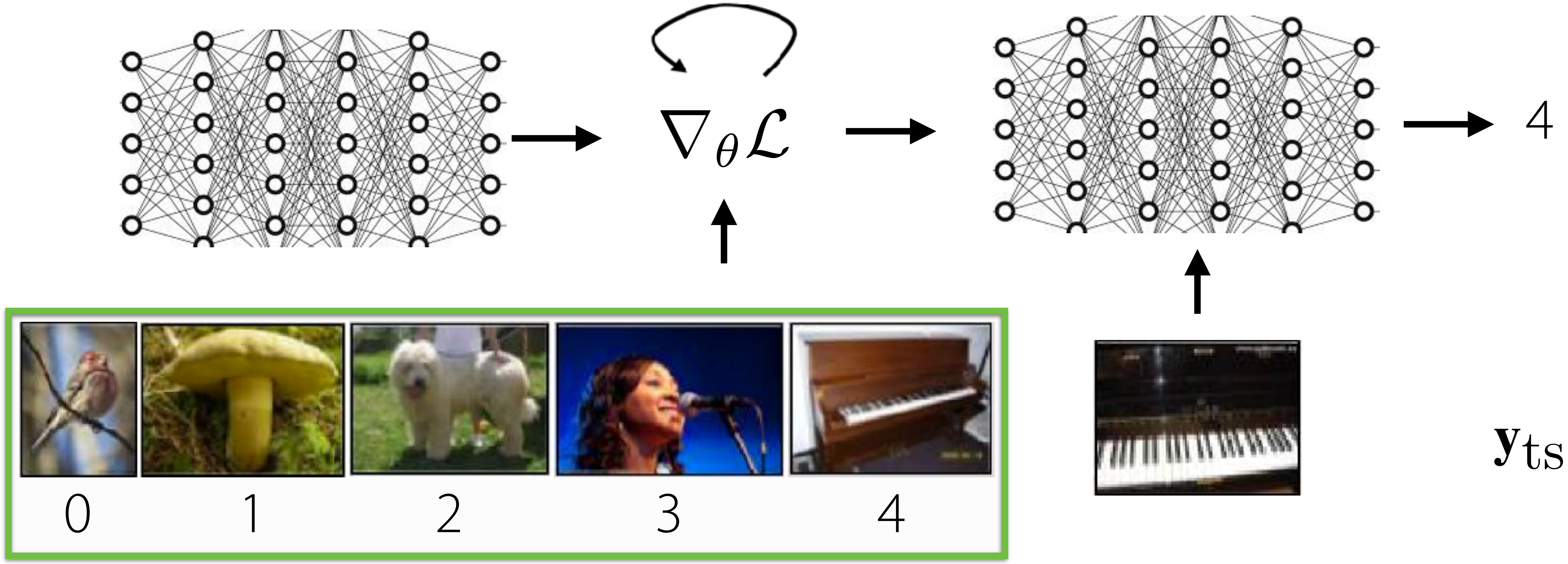
$$\mathbf{y}_{\text{ts}} = f(\mathscr{D}_{\text{tr}}, \mathbf{x}_{\text{ts}}; \theta)$$

(Hochreiter et al. '91, Santoro et al. '16, many others)

How does meta-learning work?

$\mathcal{T}_1$

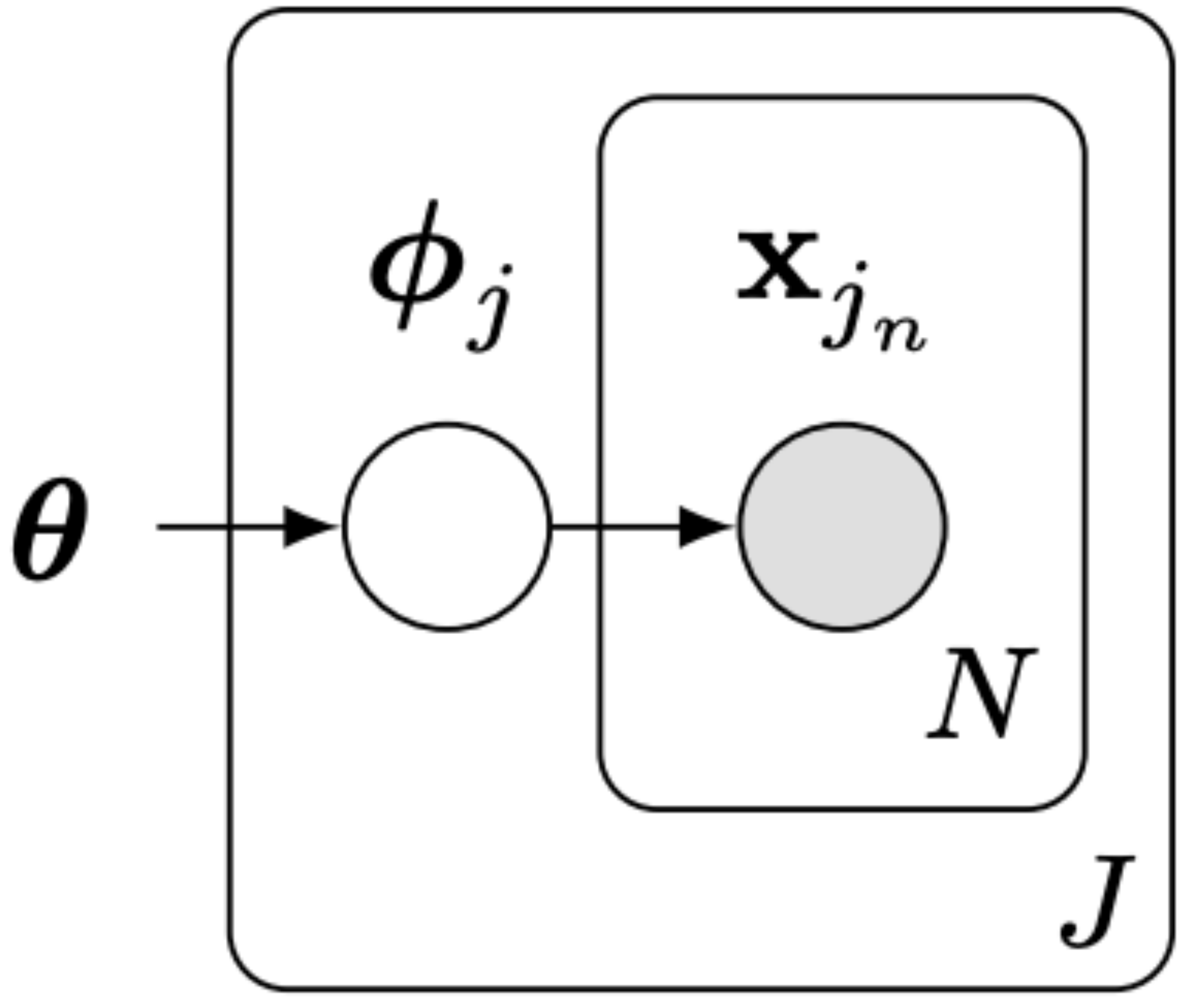
Another approach: embed optimization inside the learning process



$$\mathbf{y}_{\text{ts}} = f(\mathcal{D}_{\text{tr}}, \mathbf{x}_{\text{ts}}; \theta)$$

(Maclaurin et al. '15, Finn et al. '17, many others)

The Bayesian perspective



meta-learning <~> learning priors $p(\phi \mid \theta)$ from data

(Grant et al. '18, Gordon et al. '18, many others)

Outline

1. Brief overview of meta-learning

2. The problem: peculiar, lesser-known, yet ubiquitous

3. First steps towards a solution

How we construct tasks for meta-learning.

Randomly assign class labels to image classes for each task —> Tasks are *mutually exclusive*.

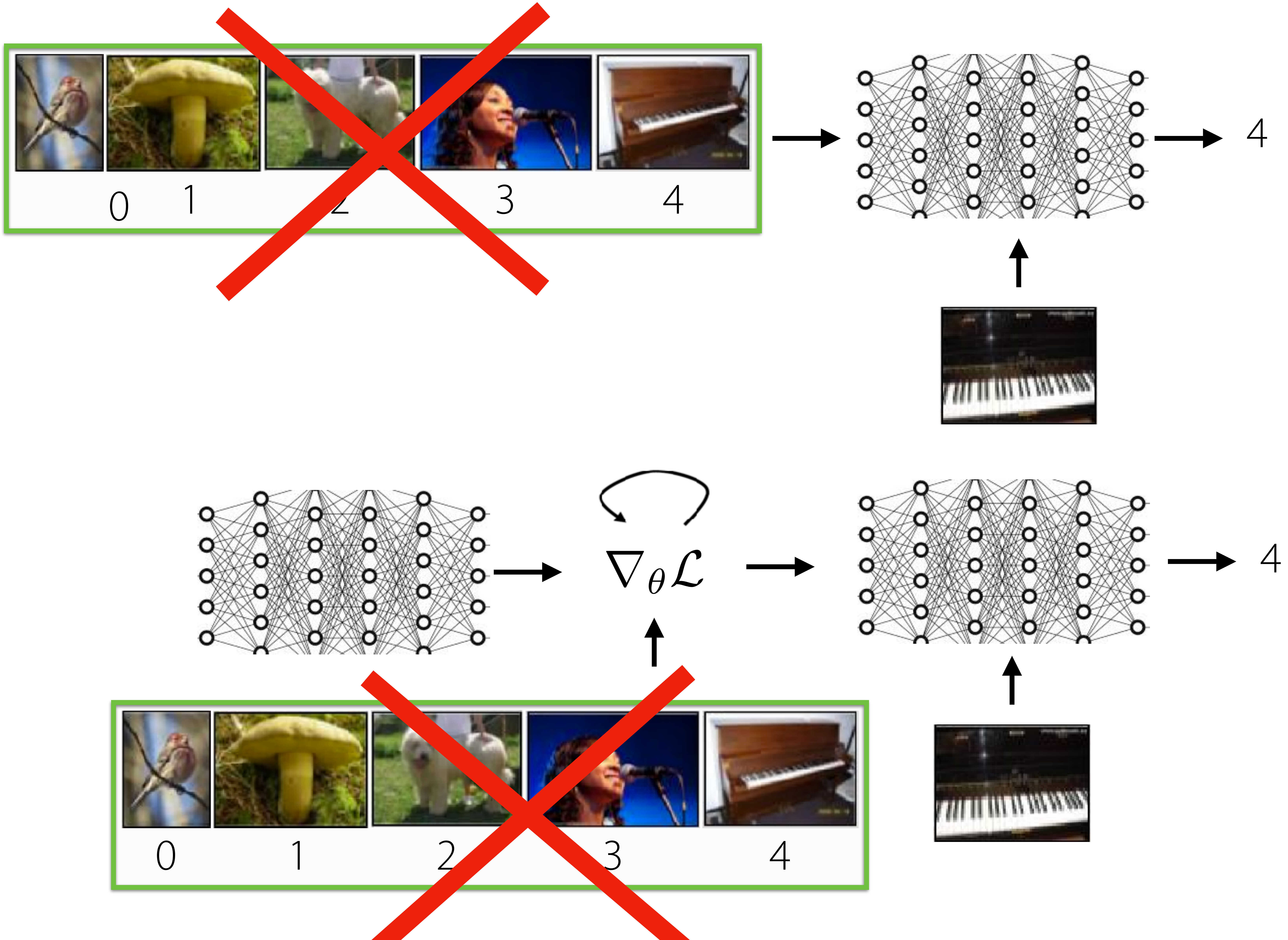
Algorithms **must** use training data to infer label ordering.

What if label order is consistent?

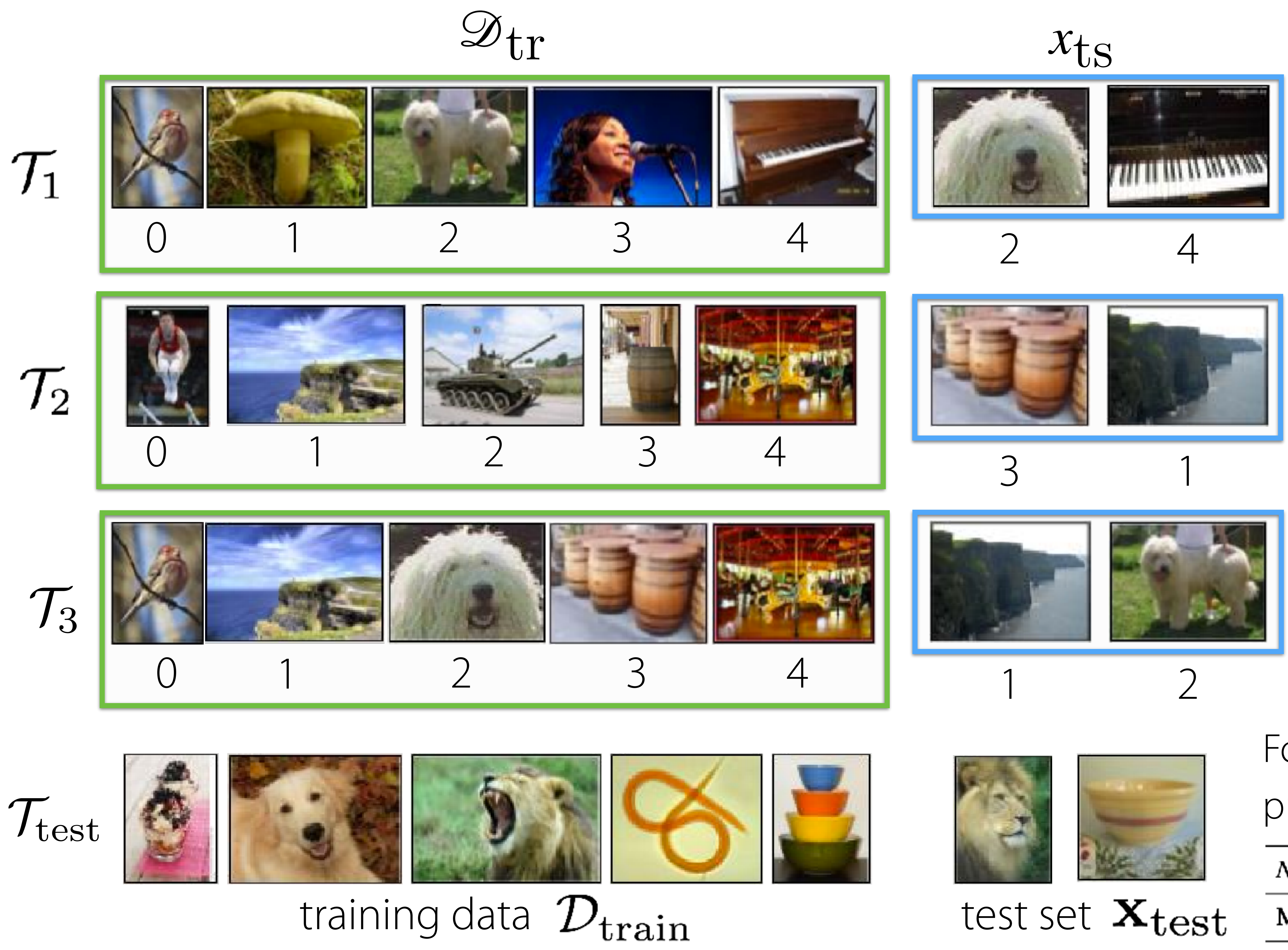
Tasks are **non-mutually exclusive**: a single function can solve all tasks.

The network can simply learn to classify inputs, irrespective of $\mathcal{D}_{\mathrm{tr}}$

The network can simply learn to classify inputs, irrespective of $\mathcal{D}_{\mathrm{tr}}$



What if label order is consistent?



For new image classes: can't make predictions w/o $\mathcal{D}_{\text{tr}}$

NME Omniglot	20-way 1-shot	20-way 5-shot
MAML	7.8 (0.2)%	50.7 (22.9)%

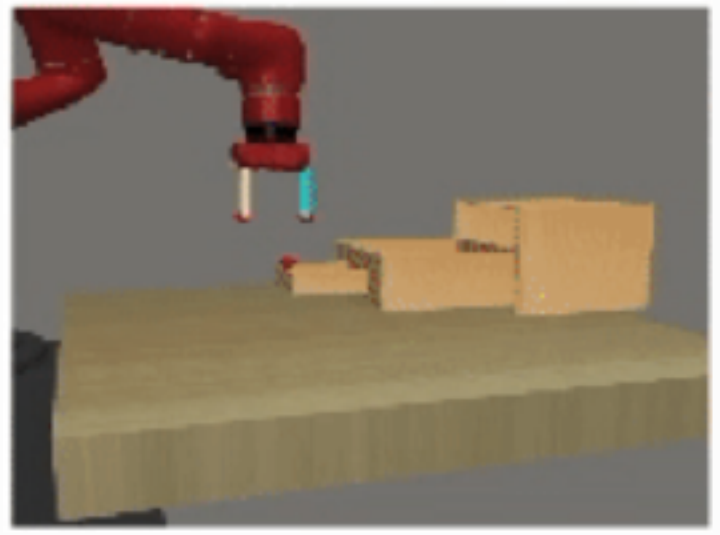
Is this a problem?

- **No**: for image classification, we can just shuffle labels*
- **No**, if we see the same image classes as training (& don't need to adapt at meta-test time)
- But, ***yes***, if we want to be able to adapt with data for new tasks.

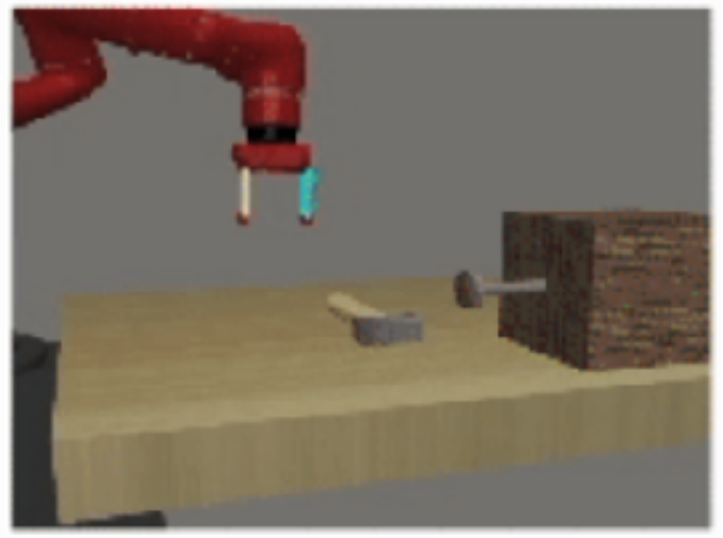
Another example

meta-training

"close drawer"

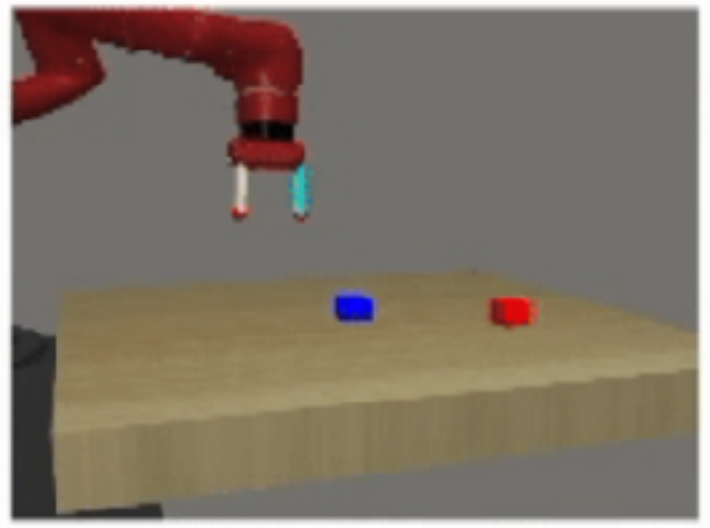
$\mathcal{T}_1$ 

"hammer"

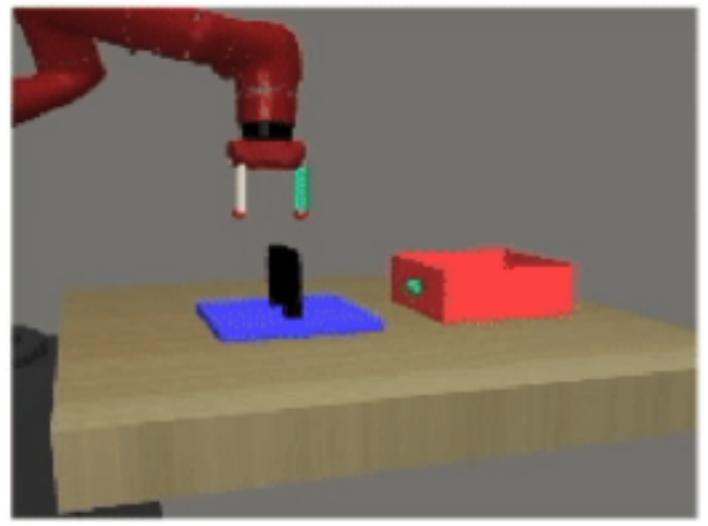
$\mathcal{T}_2$ 

…

"stack"

$\mathcal{T}_{50}$ 

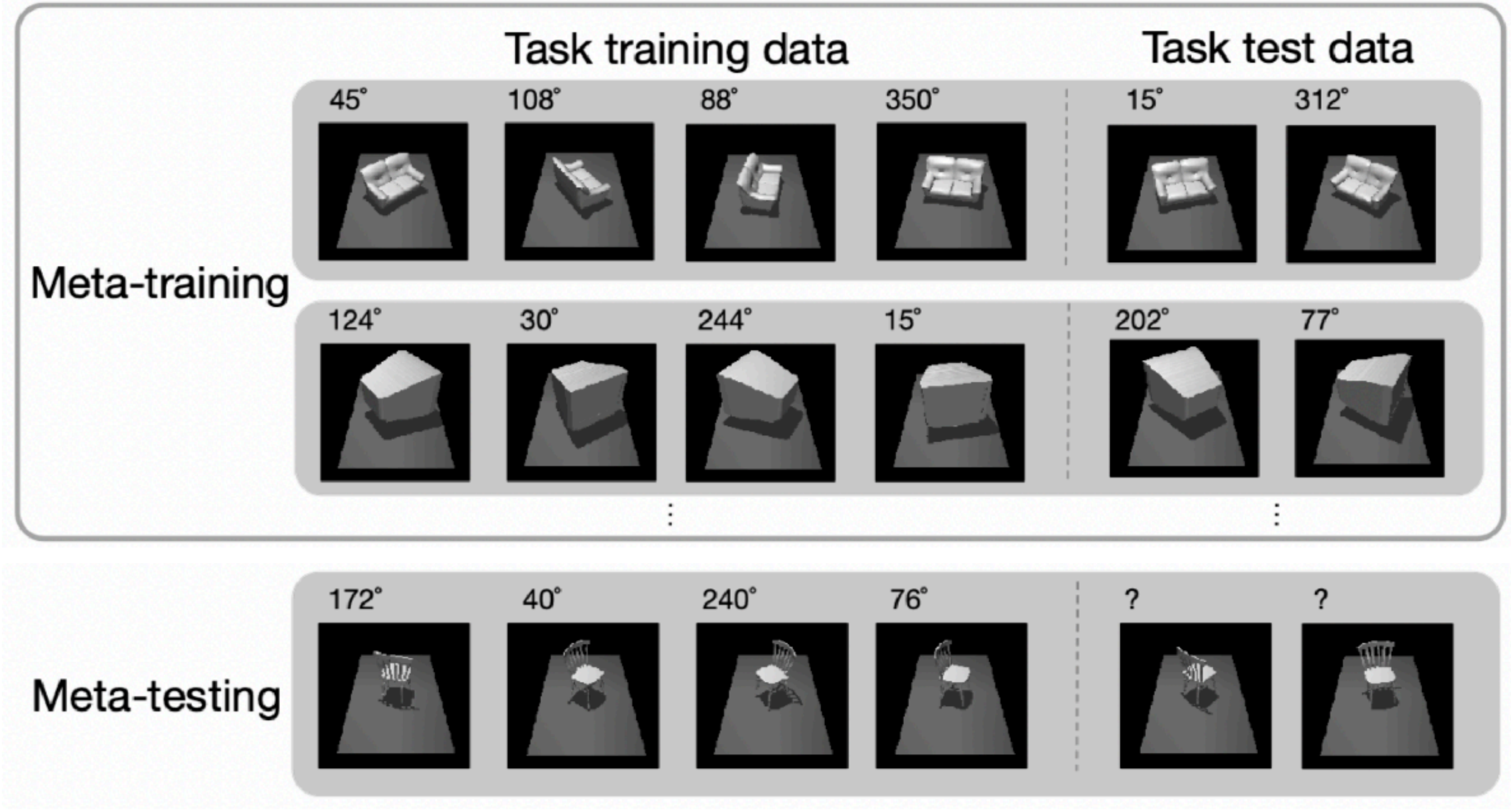
"close box"

$\mathcal{T}_{\text{test}}$ 

If you tell the robot the task goal, the robot can **ignore** the trials.

T Yu, D Quillen, Z He, R Julian, K Hausman, C Finn, S Levine. ***Meta-World***. CoRL '19

Another example



Model can memorize the canonical orientations of the training objects.

Yin, Tucker, Yuan, Levine, Finn. **Meta-Learning without Memorization**. '19

Can we do something about it?

If tasks *mutually exclusive*: single function cannot solve all tasks

(i.e. due to label shuffling, hiding information)

If tasks are *non-mutually exclusive*: single function can solve all tasks

multiple solutions to the meta-learning problem

$$y^{\text{ts}} = f_\theta(\mathcal{D}_i^{\text{tr}}, x^{\text{ts}})$$

One solution: memorize canonical pose info in $\boldsymbol{\theta}$ & ignore $\mathscr{D}_i^{\text{tr}}$

Another solution: carry no info about canonical pose in $\boldsymbol{\theta}$, acquire from $\mathscr{D}_i^{\text{tr}}$

An entire spectrum of solutions based on how **information** flows.

Suggests a potential approach: control information flow.

Yin, Tucker, Yuan, Levine, Finn. **Meta-Learning without Memorization**. '19

If tasks are *non-mutually* ***exclusive***: single function can solve all tasks

multiple solutions to the meta-learning problem

$$y^{\text{ts}} = f_\theta(\mathcal{D}_i^{\text{tr}}, x^{\text{ts}})$$

One solution: memorize canonical pose info in $\boldsymbol{\theta}$ & ignore $\mathscr{D}_i^{\text{tr}}$

Another solution: carry no info about canonical pose in $\boldsymbol{\theta}$, acquire from $\mathscr{D}_i^{\text{tr}}$

An entire spectrum of solutions based on how **information** flows.

Meta-regularization one option: $\max I(\hat{\mathbf{y}}_{\text{ts}}, \mathscr{D}_{\text{tr}} \mid \mathbf{x}_{\text{ts}})$

minimize meta-training loss + information in $\boldsymbol{\theta}$

$$\mathscr{L}(\theta, \mathscr{D}_{meta-train}) + \beta D_{KL}(q(\theta; \theta_\mu, \theta_\sigma) \| p(\theta))$$

Places precedence on using information from $\mathscr{D}_{\text{tr}}$ over storing info in $\boldsymbol{\theta}$.

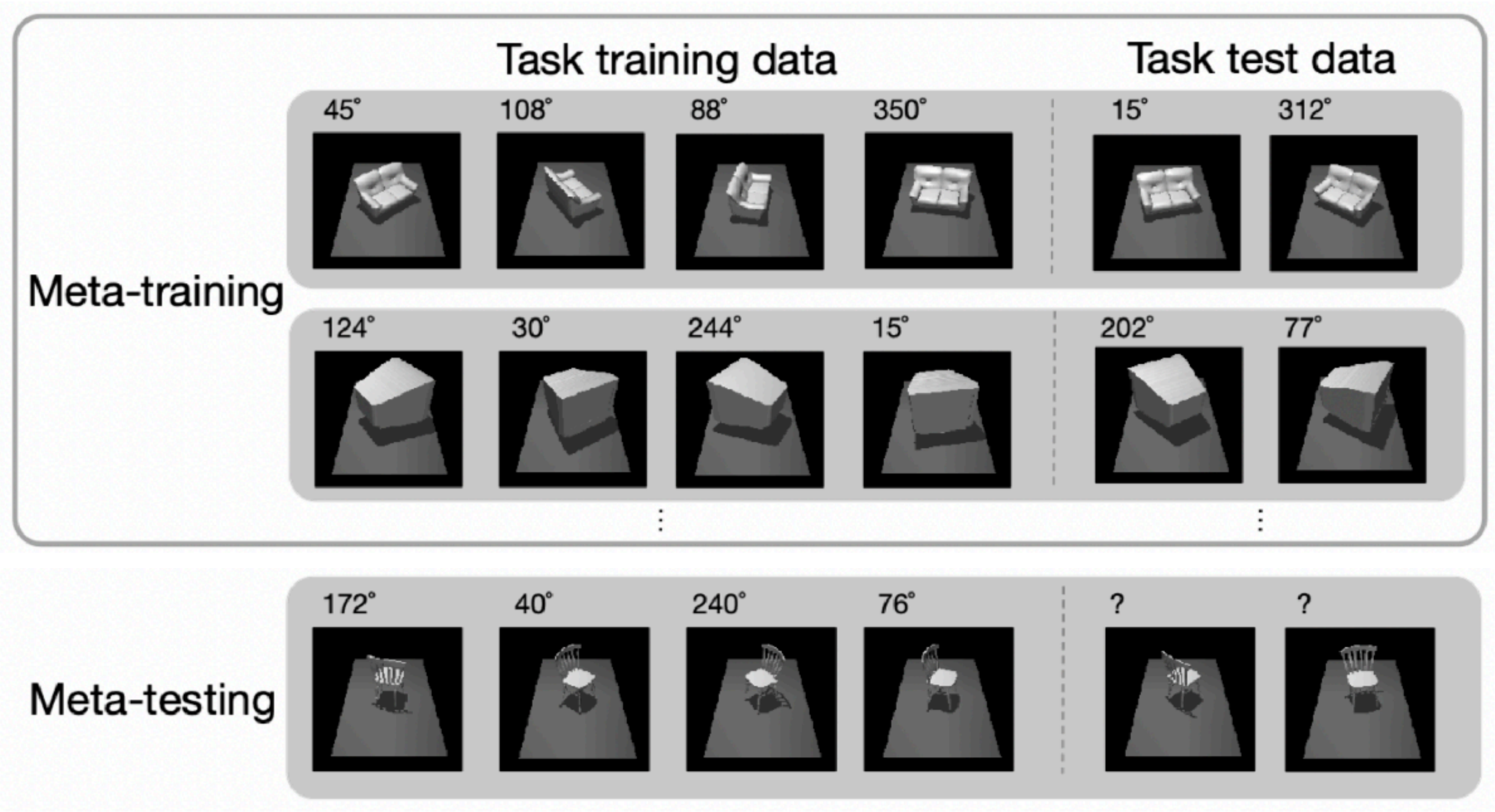
Can combine with your favorite meta-learning algorithm.

Yin, Tucker, Yuan, Levine, Finn. **Meta-Learning without Memorization.** '19

Omniglot without label shuffling: "non-mutually-exclusive" Omniglot

NME Omniglot	20-way 1-shot	20-way 5-shot
MAML	7.8 (0.2)%	50.7 (22.9)%
TAML	9.6 (2.3)%	67.9 (2.3)%
MR-MAML (W) (ours)	**83.3 (0.8)**%	**94.1 (0.1)**%

On **pose prediction** task:



Method	MAML	MR-MAML(W) (ours)	CNP	MR-CNP(W) (ours)
MSE	5.39 (1.31)	**2.26 (0.09)**	8.48 (0.12)	2.89 (0.18)

(and it's not just as simple as standard regularization)

CNP	CNP + Weight Decay	CNP + BbB	MR-CNP (W) (ours)
8.48 (0.12)	6.86 (0.27)	7.73 (0.82)	**2.89 (0.18)**

TAML: Jamal & Qi. **Task-Agnostic Meta-Learning for Few-Shot Learning**. CVPR '19

Yin, Tucker, Yuan, Levine, Finn. **Meta-Learning without Memorization**. '19

Does meta-regularization lead to better generalization?

Let $P(\theta)$ be an arbitrary distribution over θ that doesn't depend on the meta-training data.

(e.g. $P(\theta) = \mathcal{N}(\theta; \mathbf{0}, \mathbf{I})$)

For MAML, with probability at least $1 - \delta$,

$$\underbrace{er(\theta_\mu, \theta_\sigma)}_{\text{generalization error}} \leq \underbrace{\frac{1}{n}\sum_{i=1}^{n} \hat{er}(\theta_\mu, \theta_\sigma, \mathcal{D}_i, \mathcal{D}_i^*)}_{\text{error on the meta-training set}} + \left(\sqrt{\frac{1}{2(K-1)}} + \sqrt{\frac{1}{2(n-1)}}\right)\sqrt{\underbrace{D_{KL}(\mathcal{N}(\theta; \theta_\mu, \theta_\sigma)\|P)}_{\text{meta-regularization}} + \log\frac{n(K+1)}{\delta}},$$

$\forall \theta_\mu, \theta_\sigma$

With a Taylor expansion of the RHS + a particular value of β —> *recover the MR MAML objective*.

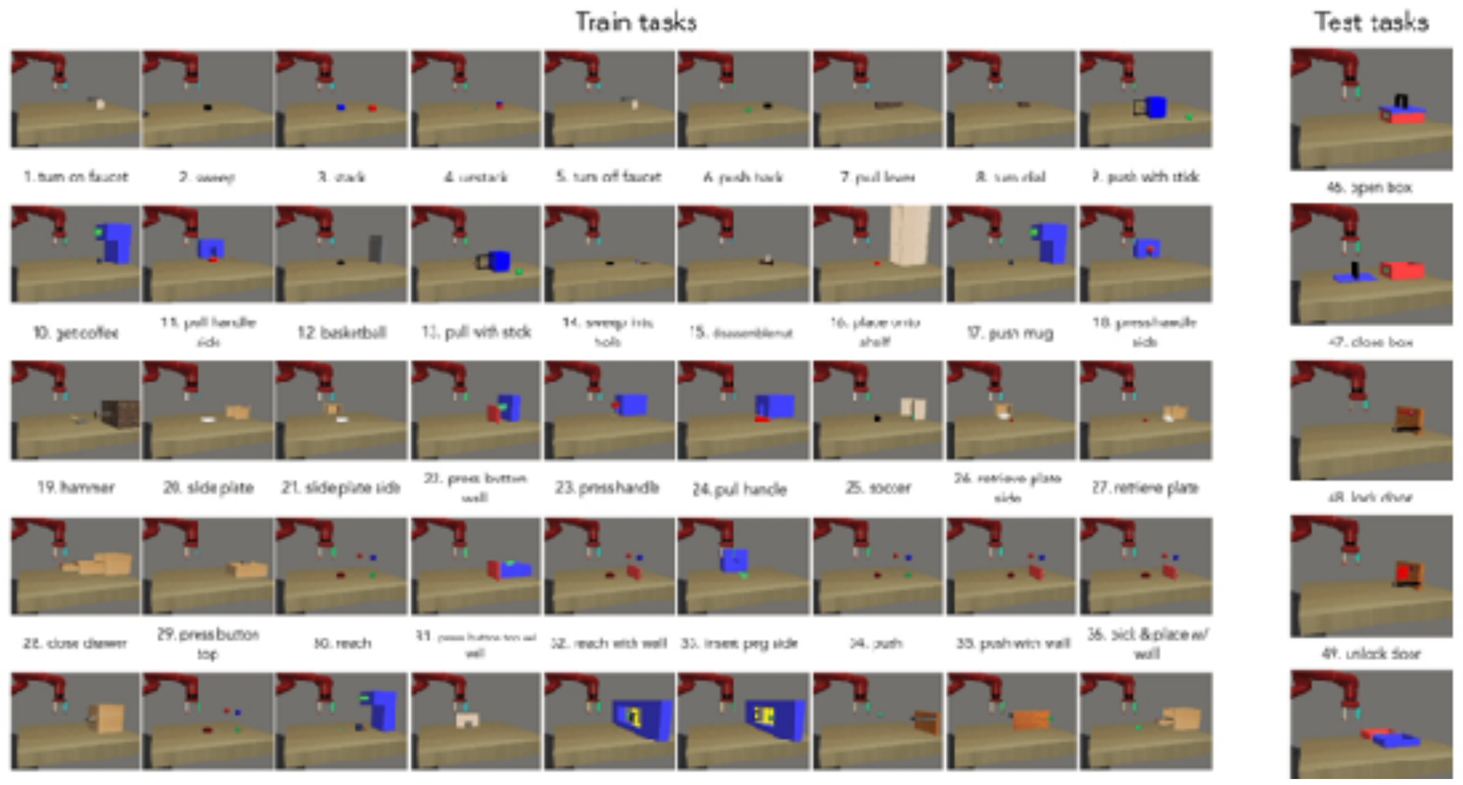
Proof: draws heavily on Amit & Meier '18

Yin, Tucker, Yuan, Levine, Finn. **Meta-Learning without Memorization**. '19

Want to Learn More?

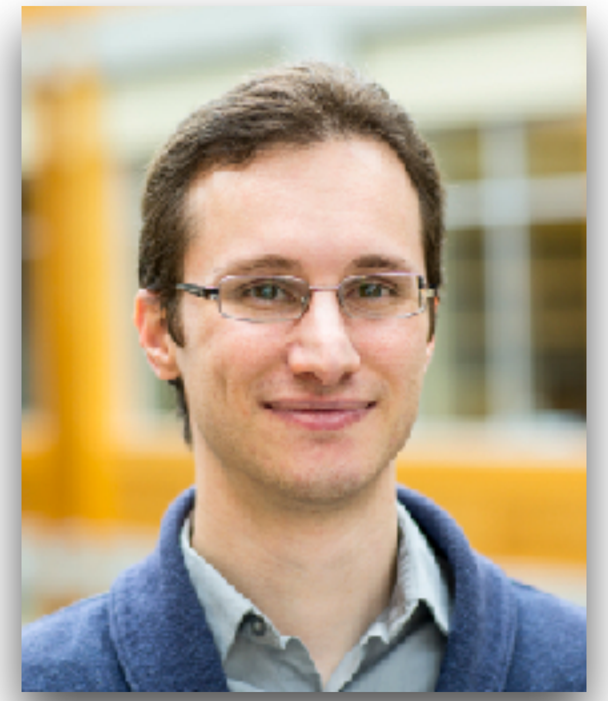
CS330: Deep Multi-Task & Meta-Learning
Lecture videos coming out soon!

Working on Meta-RL?



Try out the Meta-World benchmark

Collaborators



T Yu, D Quillen, Z He, R Julian, K Hausman, C Finn, S Levine. ***Meta-World***. CoRL '19

Yin, Tucker, Yuan, Levine, Finn. **Meta-Learning without Memorization**. '19