Advanced Meta-Learning: Task Construction

CS 330

Logistics

Today:

Homework 1 due, Homework 2 out

Project survey due (**non-binding**, for assigning **mentor/PoC**)

Tomorrow at 4 pm:

Tutorial on variational inference

Question of the Day

How should tasks be defined for good meta-learning performance?

Plan for Today

Brief Recap of Meta-Learning & Task Construction

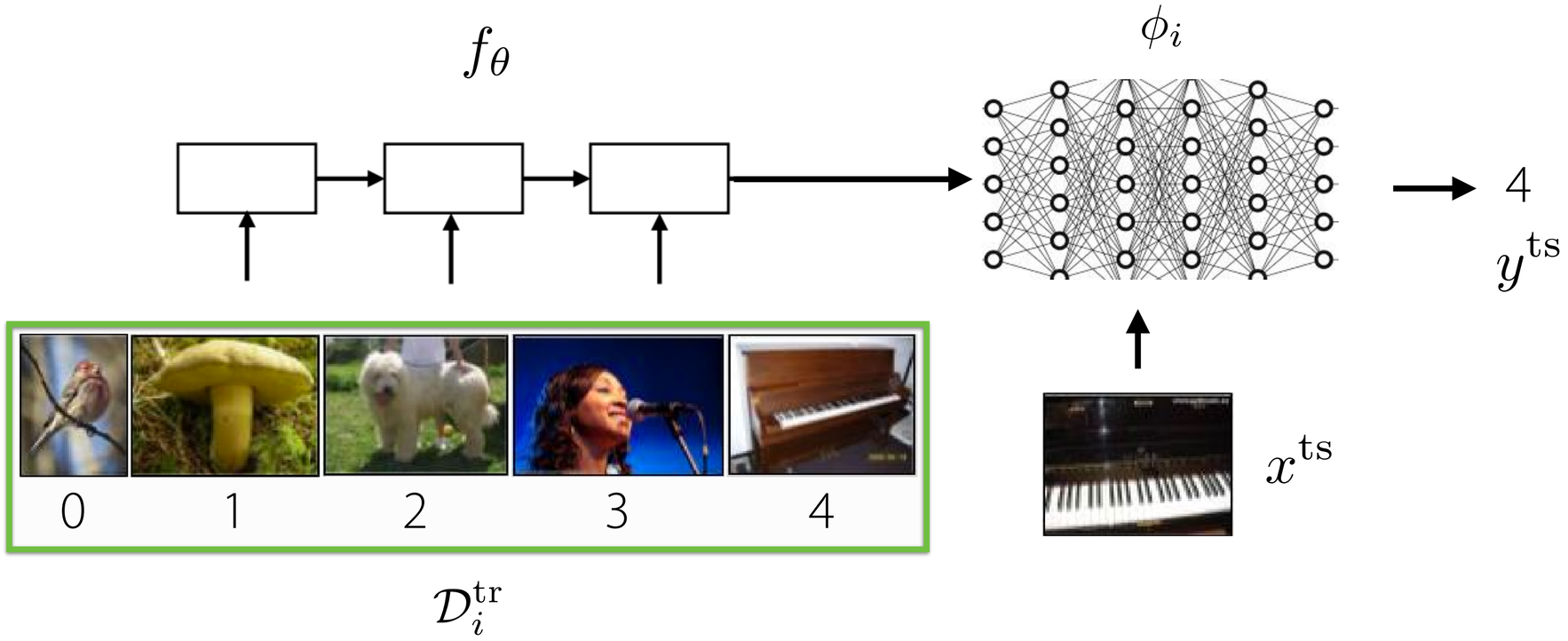
Memorization in Meta-Learning
- When it arises
- Potential solutions

Meta-Learning without Tasks Provided
- Unsupervised Meta-Learning
- Meta-Learning from Unsegmented Task Stream (time permitting)

🚨 **Disclaimer** 🚨**:** These topics are at the bleeding edge of research.

Goals for by the end of lecture:
- Understand when & how memorization in meta-learning may occur
- Understand techniques for constructing tasks automatically

Recap: Black-Box Meta-Learning

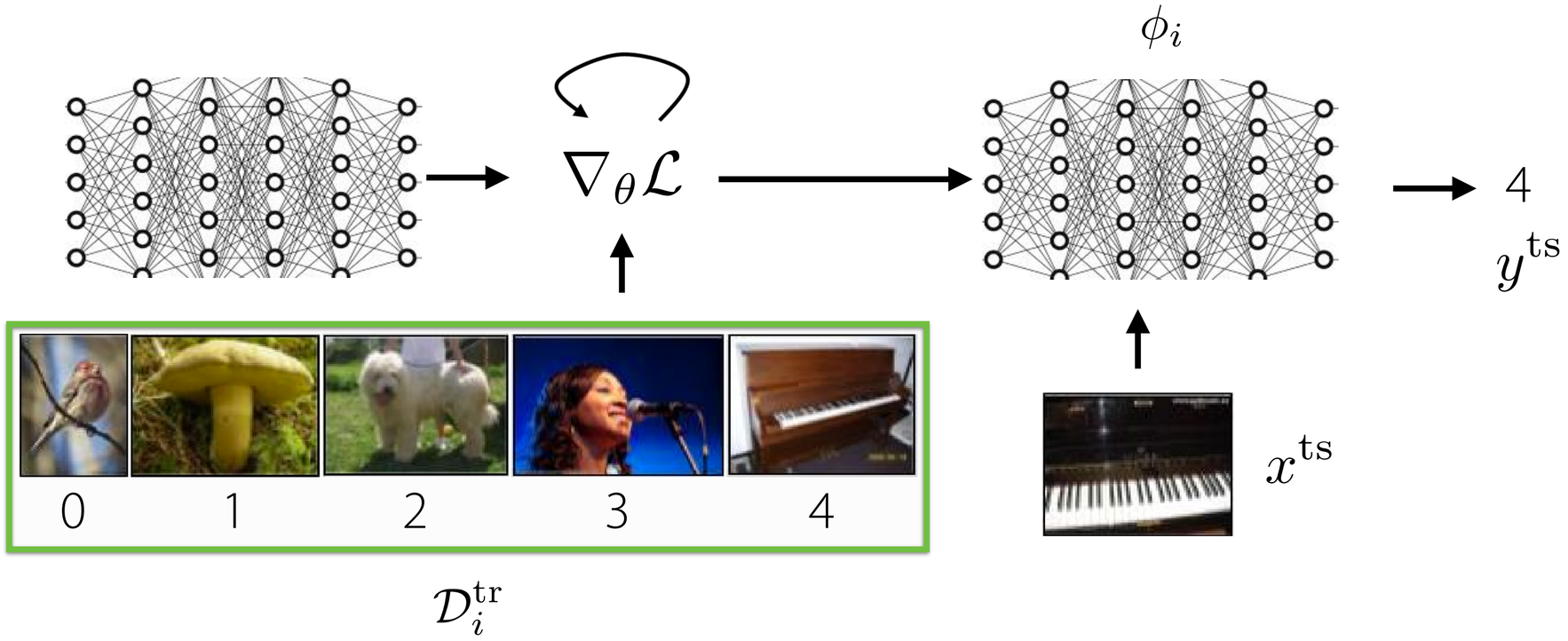


Key idea: parametrize learner as a neural network

+ **expressive**

- **challenging optimization** problem

Recap: Optimization-Based Meta-Learning

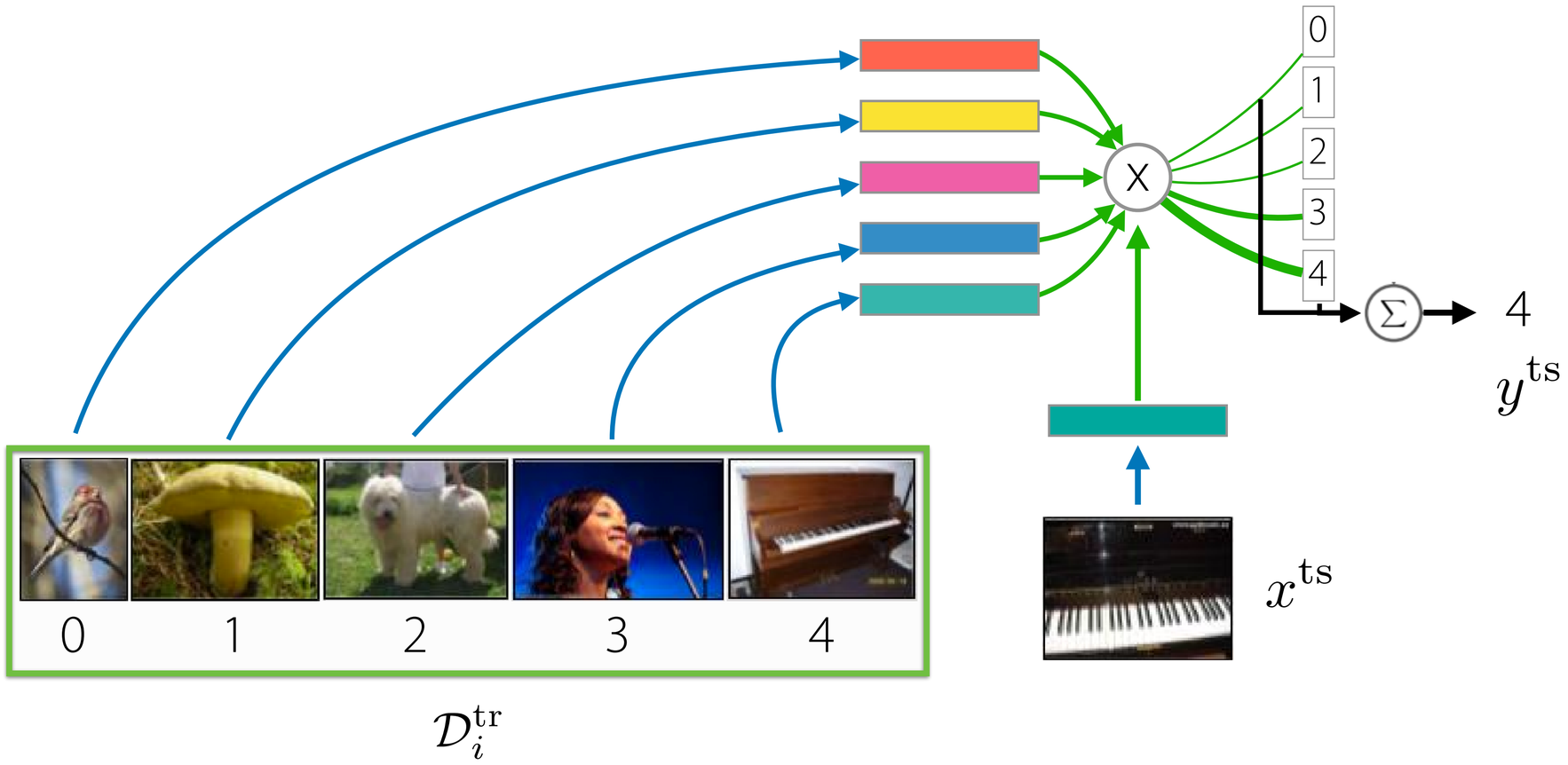


Key idea: embed optimization inside the inner learning process

+ **structure** of **optimization** embedded into meta-learner

- typically requires **second-order optimization**

Recap: Non-Parametric Meta-Learning



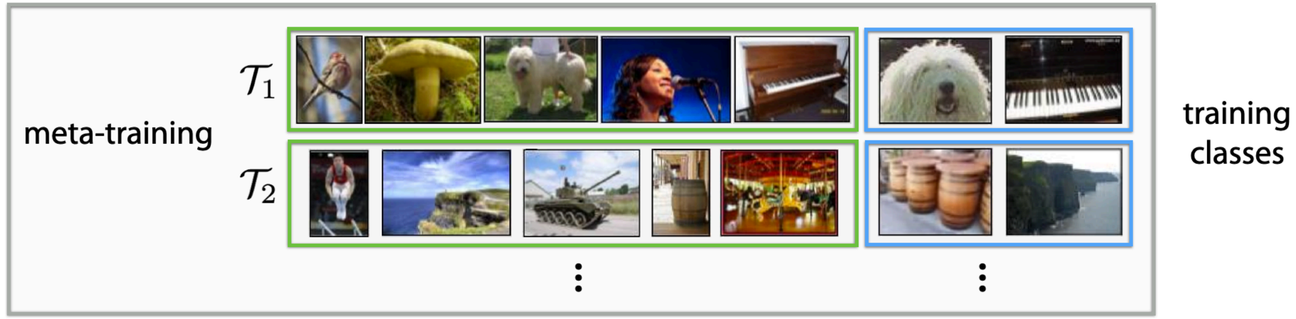
Key idea: *non-parametric learner* with *parametric* embedding / distance
(e.g. kNN to examples/prototypes)

+ easy to optimize, computationally fast

- largely restricted to classification

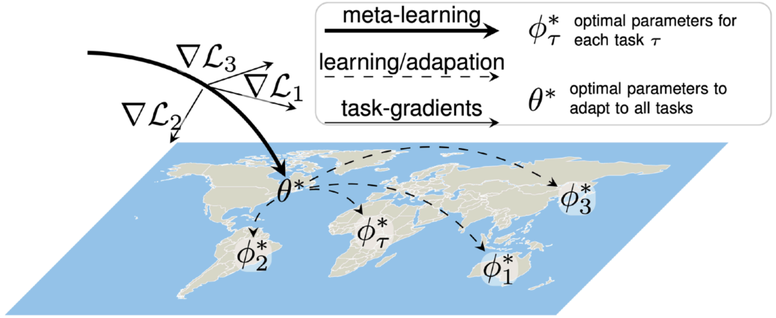
Recap: Task Construction Techniques

For N-way image classification



Use **labeled images** from prior classes

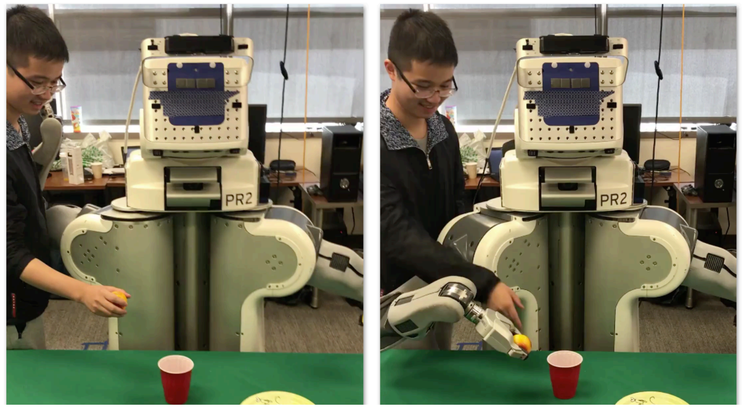
For adapting to regional differences



Rußwurm et al. Meta-Learning for Few-Shot Land Cover Classification. CVPR 2020 EarthVision Workshop

Use **labeled images** from prior regions

For few-shot imitation learning



Yu et al. One-Shot Imitation Learning from Observing Humans. RSS 2018

Use **demonstrations** for prior tasks

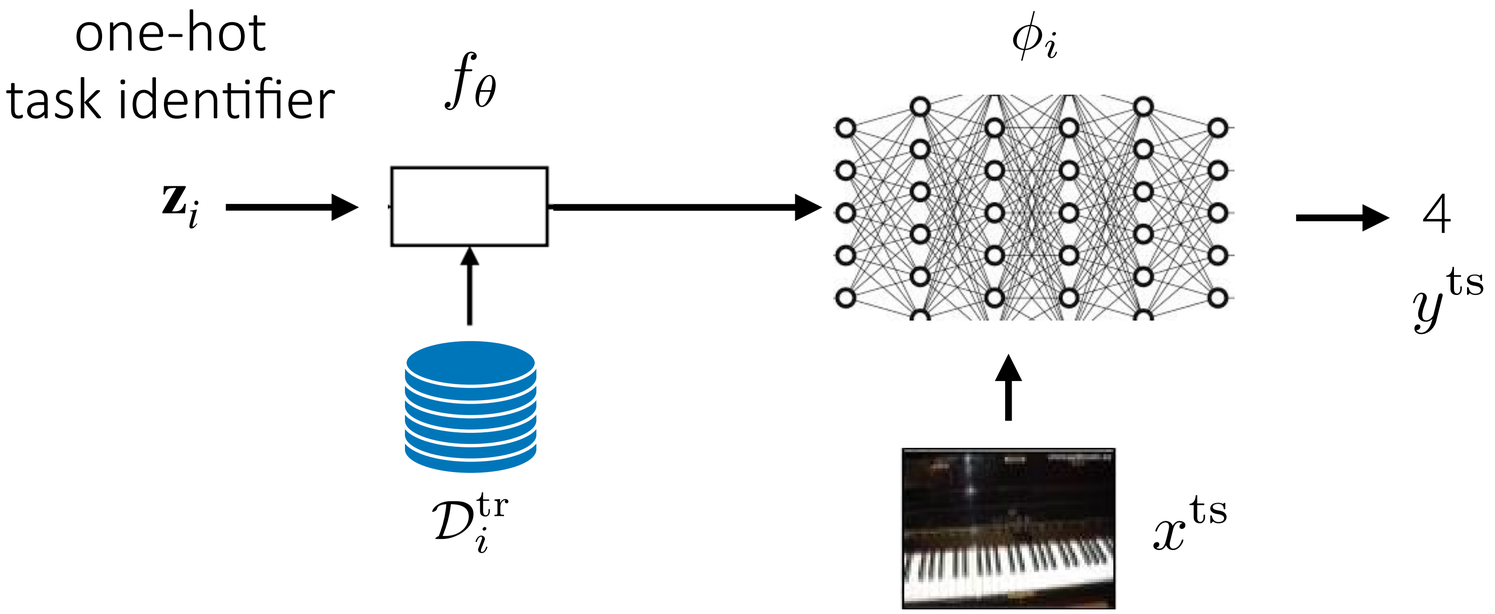
Plan for Today

Brief Recap of Meta-Learning & Task Construction

Memorization in Meta-Learning
- When it arises
- Potential solutions

Meta-Learning without Tasks Provided
- Unsupervised Meta-Learning
- Meta-Learning from Unsegmented Task Stream (time permitting)

Thought Exercise #1



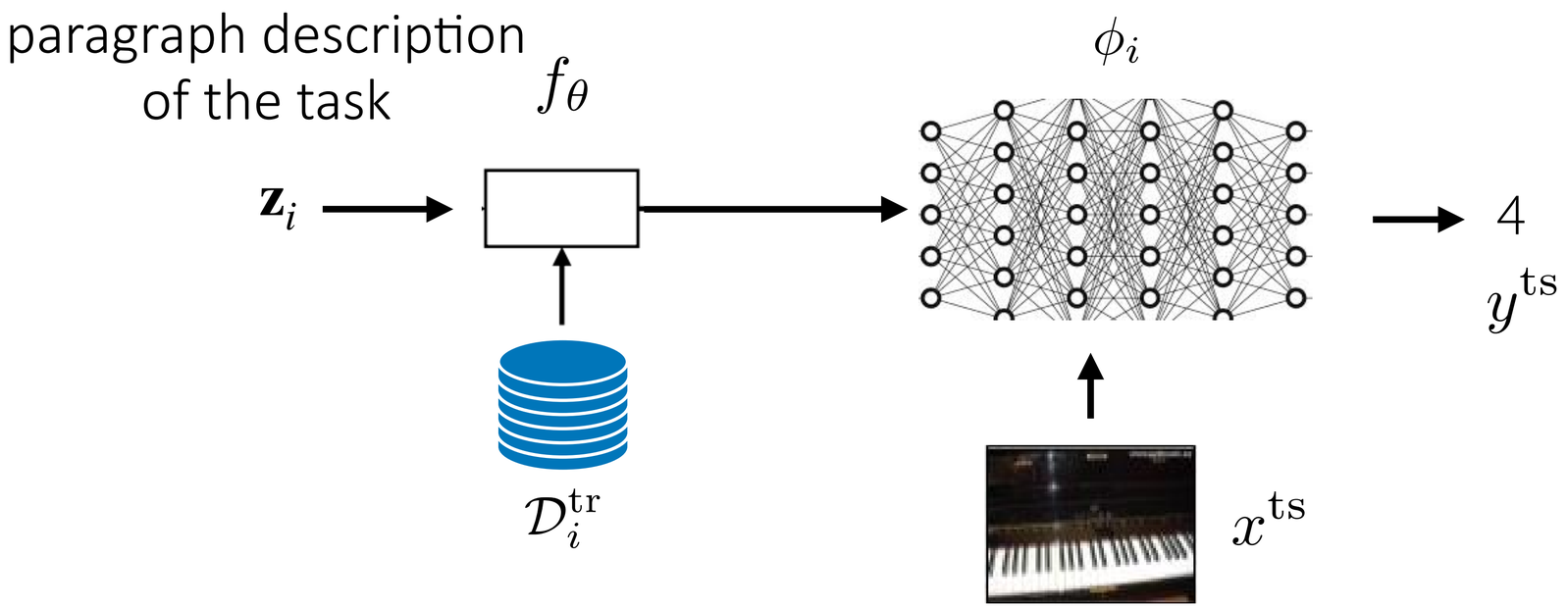
Question: What happens during <u>meta-training</u> if you pass in D_i^{tr} ***and*** the task identifier?

If it is difficult to learn from the data, the model will learn rely on $\mathbf{z}_i$.

Question: What happens at <u>meta-test time</u> if you pass in D_j^{tr} ***and*** the task identifier for a <u>new task</u>?

It won't generalize to the new task.

Thought Exercise #2



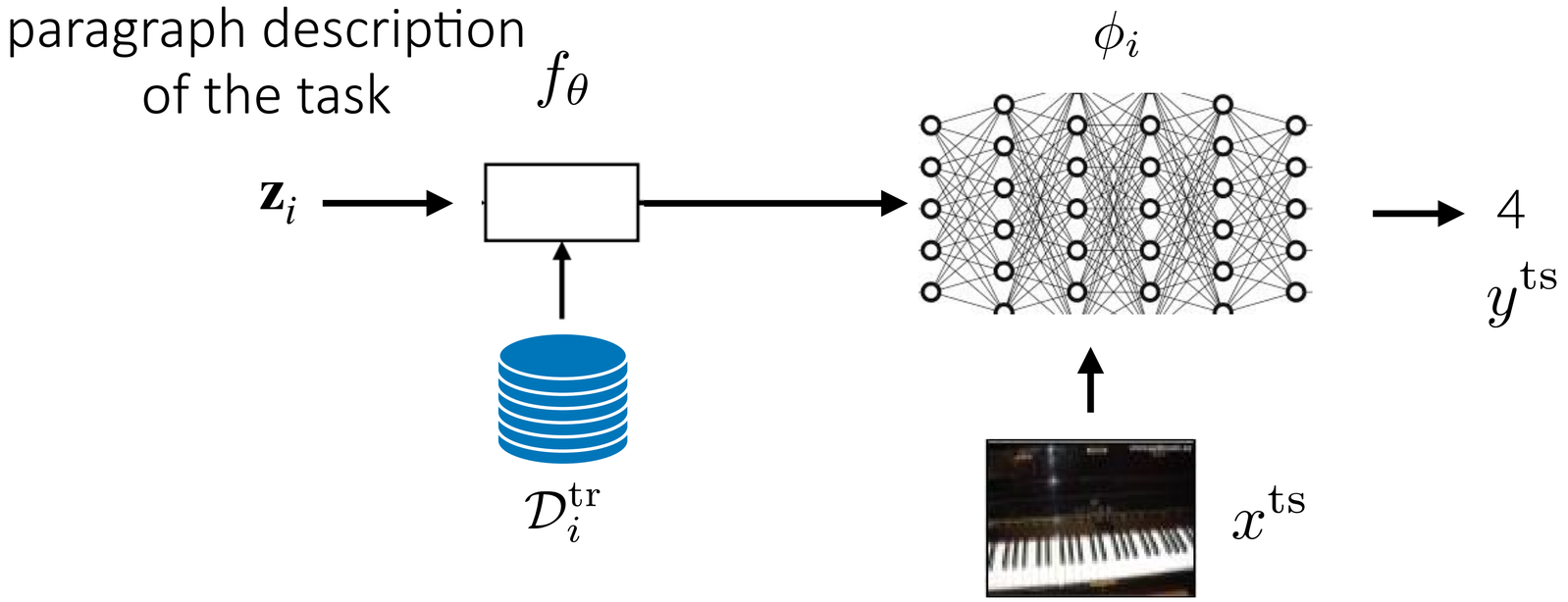
Question: What happens during <u>meta-training</u> if you pass in D_i^{tr} ***and*** the task identifier?

It depends on whether using the **description** or the **data** is simpler.

Question: What happens at <u>meta-test time</u> if you pass in D_j^{tr} ***and*** the task identifier for a <u>new task</u>?

It depends on what it learns to use during meta-training.

Thought Exercise #2



Question: What happe ***and*** the task identifier?

It depends o ta is simpler.

Key problem: Model can minimize meta-training loss without looking at D_i^{tr}

Question: What happens at meta-test time if you pass in D_j^{tr} ***and*** the task identifier for a new task?

It depends on what it learns to use during meta-training.

How we construct tasks for meta-learning.

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

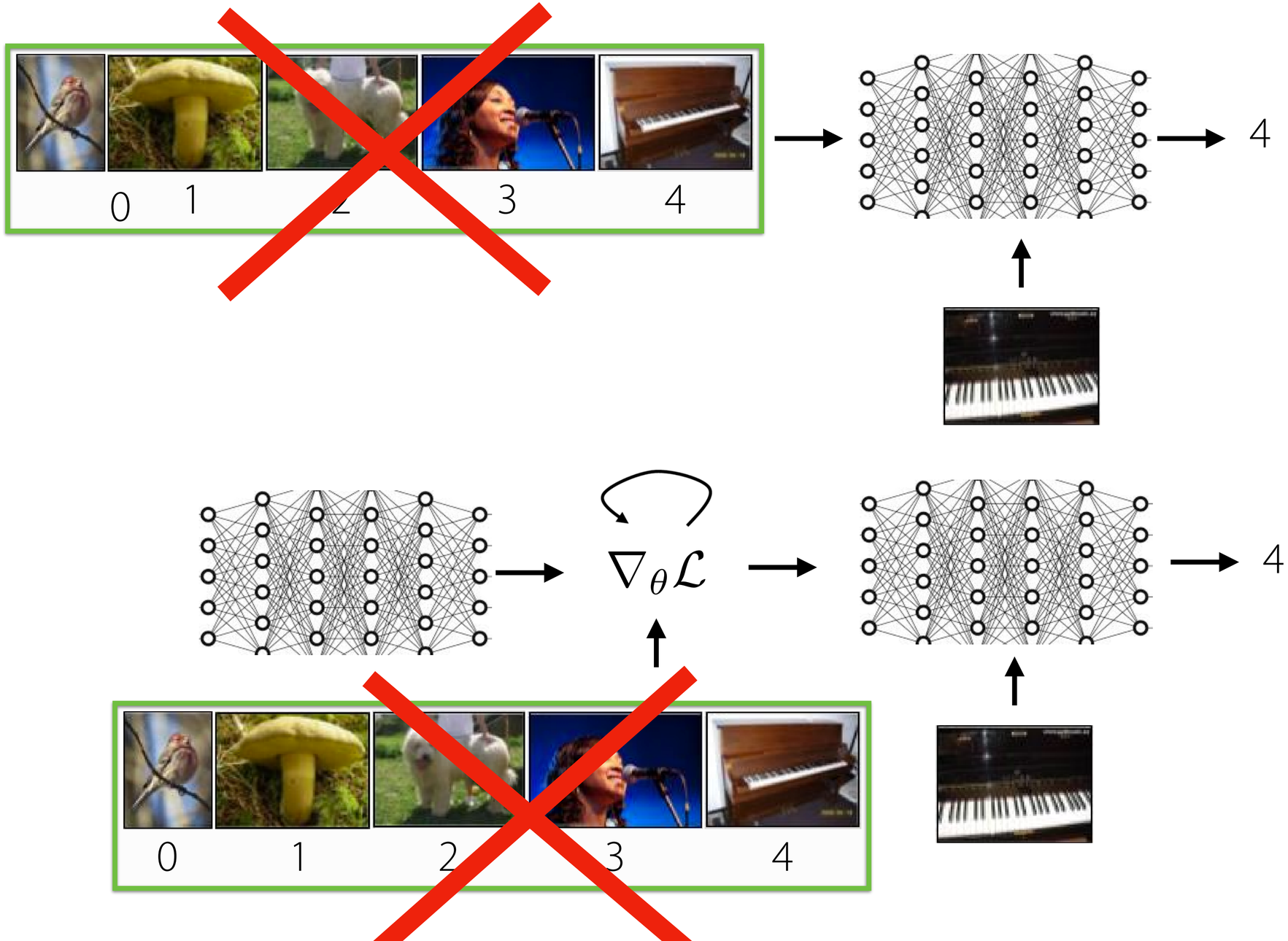
Algorithms **must** use training data to infer label ordering.

Thought Exercise #3: What if label assignment is consistent across tasks?

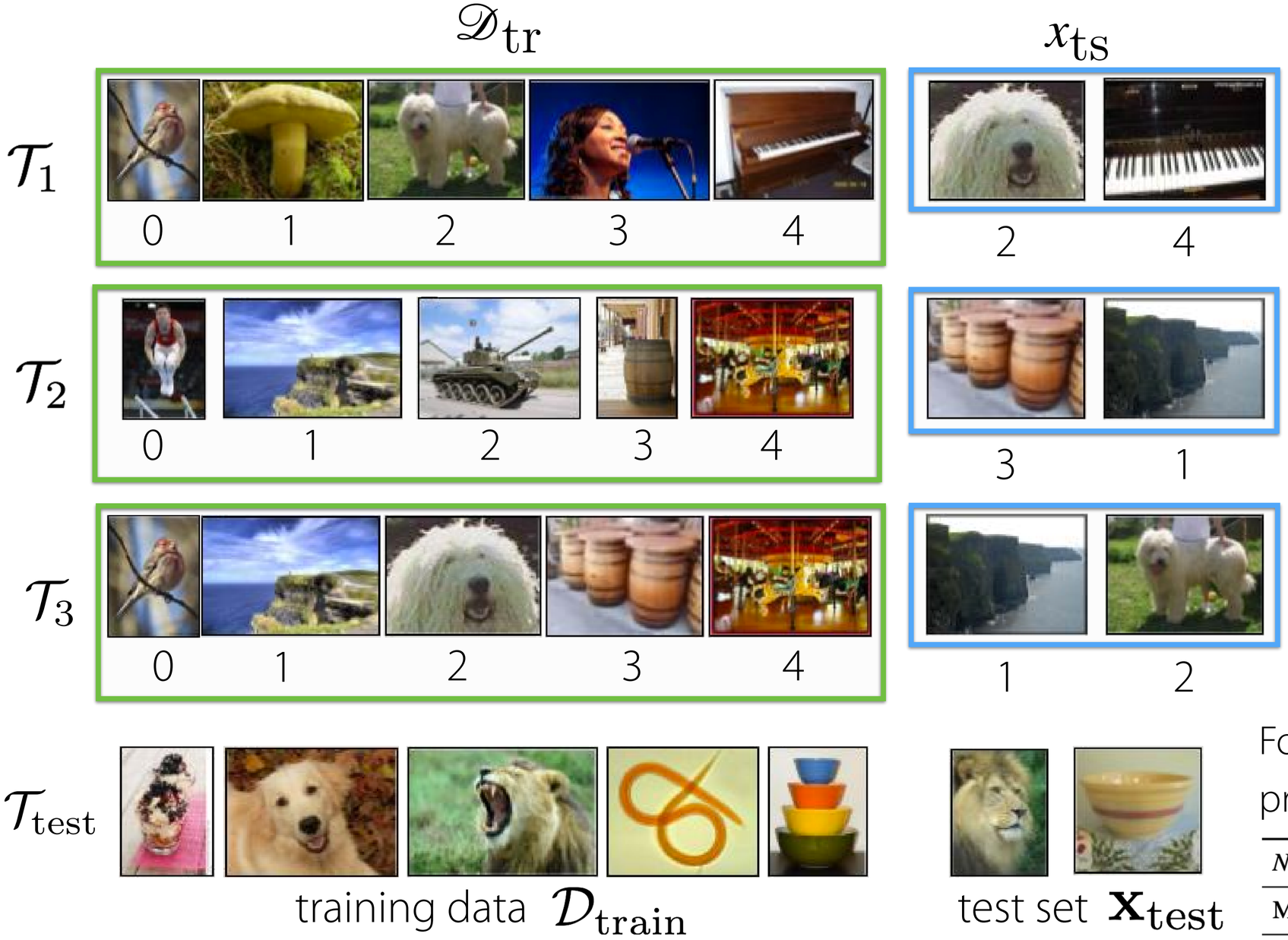
Tasks are **non-mutually exclusive**: a single function can solve all tasks.

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

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



What if label order is consistent?



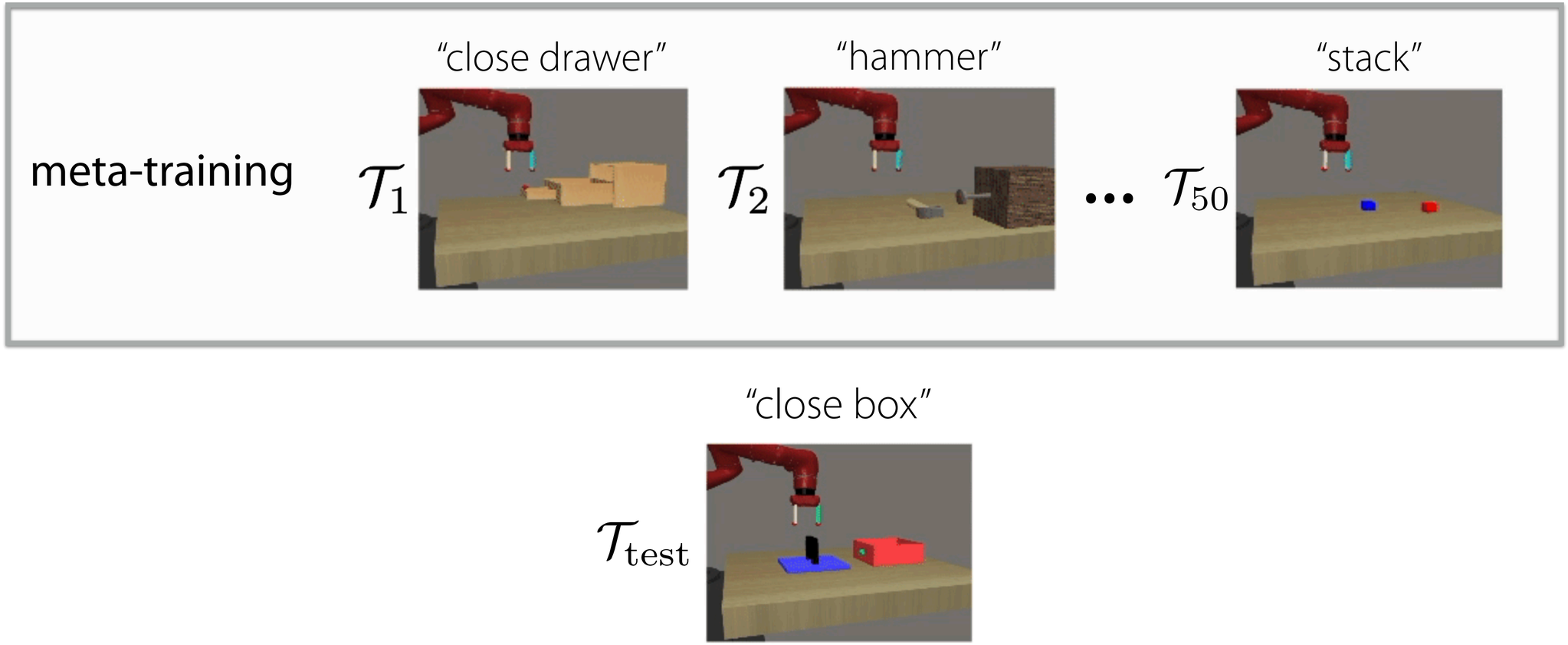
For new image classes: can't make predictions w/o $\mathscr{D}_{\text{tr}}$

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.

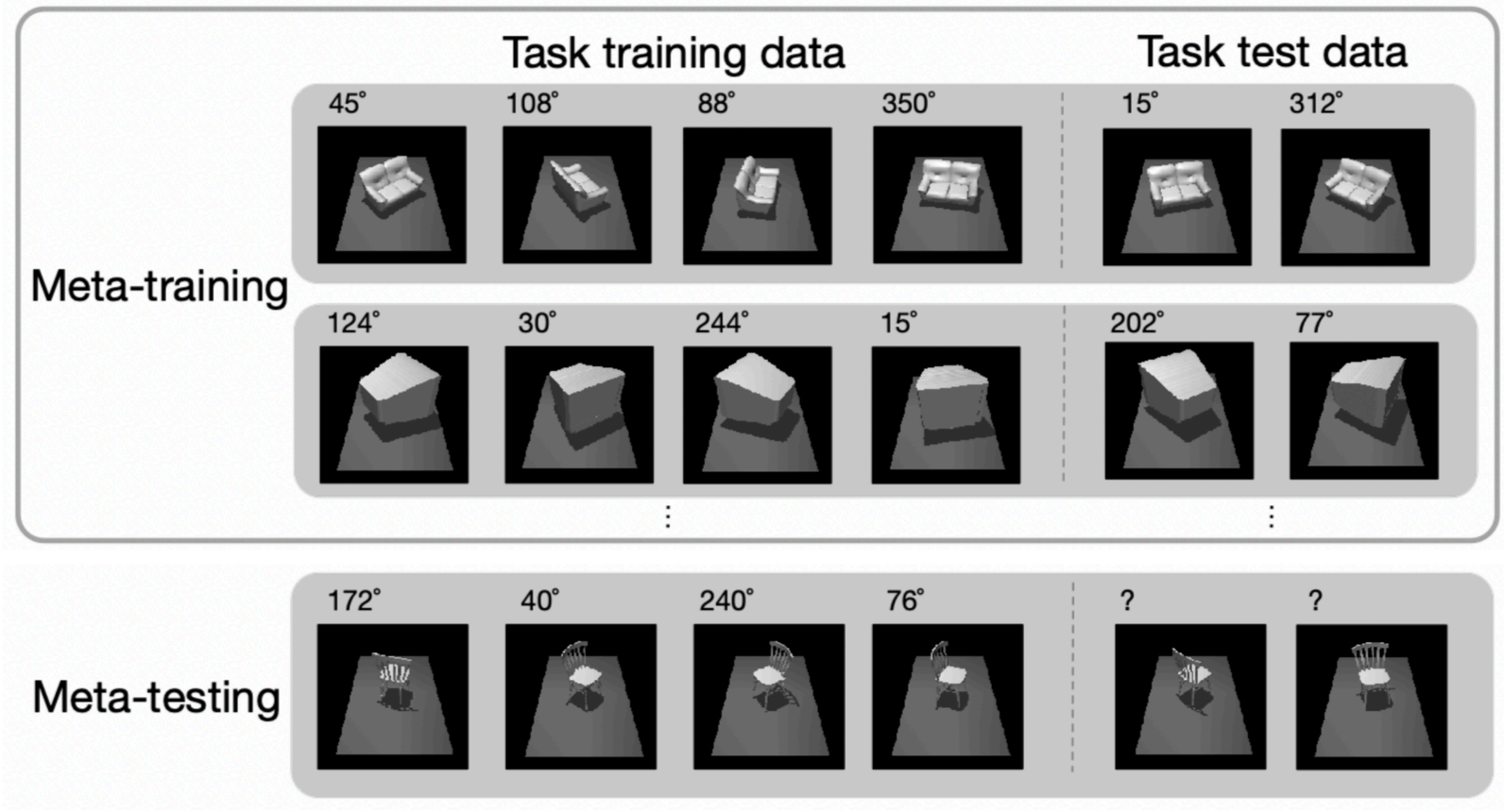
Another example



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**. ICLR '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

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

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

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

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**. ICLR '19

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

multiple solutions to the meta-learning problem

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

One solution: memorize canonical pose info in θ & ignore $\mathcal{D}_i^{\mathrm{tr}}$

Another solution: carry no info about canonical pose in θ, acquire from $\mathcal{D}_i^{\mathrm{tr}}$

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

Meta-regularization one option: $\max I(\hat{\mathbf{y}}_{\mathrm{ts}}, \mathcal{D}_{\mathrm{tr}} \,|\, \mathbf{x}_{\mathrm{ts}})$

minimize meta-training loss + information in θ

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

Places precedence on using information from $\mathcal{D}_{\mathrm{tr}}$ over storing info in θ.

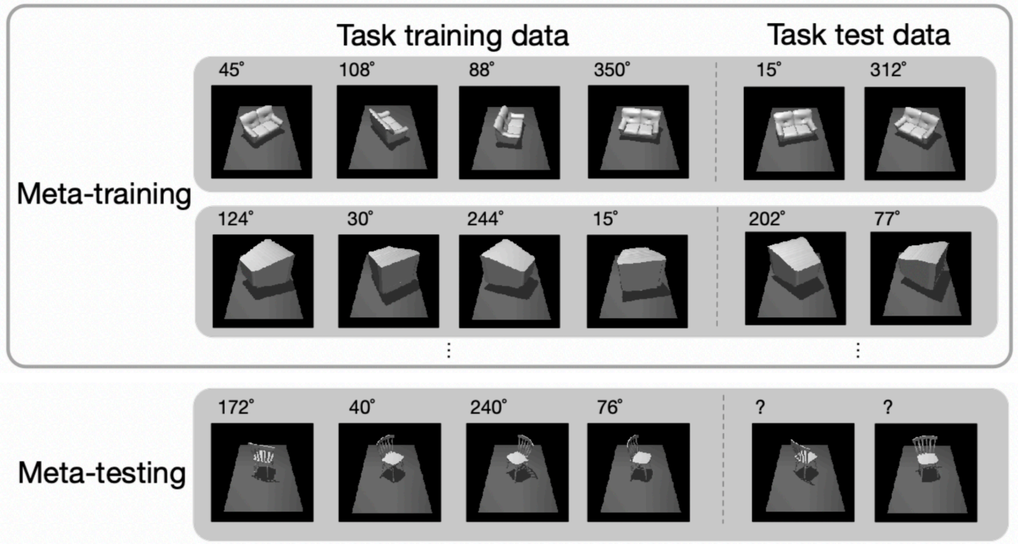
Can combine with your favorite meta-learning algorithm.

Yin, Tucker, Yuan, Levine, Finn. **Meta-Learning without Memorization**. ICLR '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**. ICLR '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 $1 - \delta$,

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

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

Proof: draws heavily on Amit & Meier '18

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

Summary of Memorization Problem

meta-learning	**standard supervised learning**
meta overfitting	standard overfitting
memorize training functions f_i	memorize training datapoints (x_i, y_i)
corresponding to tasks in your meta-training dataset	in your training dataset
meta regularization	standard regularization
controls information flow	regularize hypothesis class
regularizes description length of meta-parameters	(though not always for DNNs)

Plan for Today

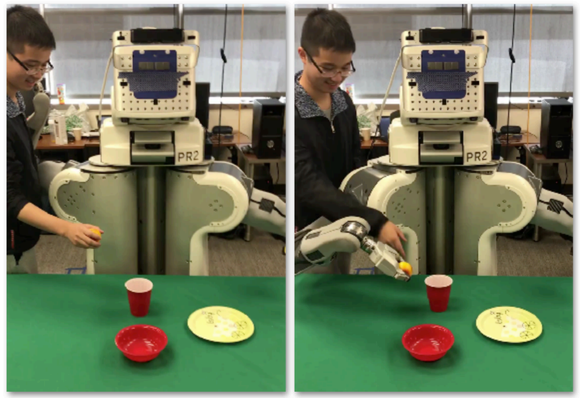
Brief Recap of Meta-Learning & Task Construction

Memorization in Meta-Learning
- When it arises
- Potential solutions

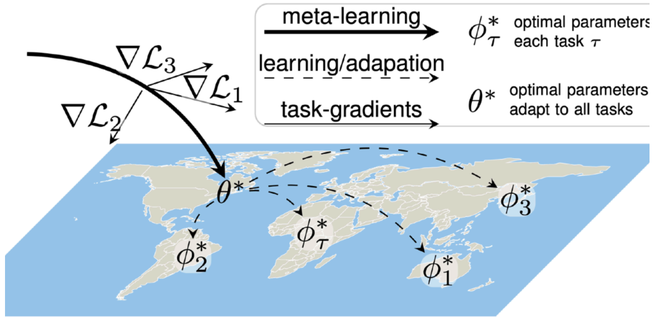
Meta-Learning without Tasks
- Unsupervised Meta-Learning
- Meta-Learning from Unsegmented Task Stream (time permitting)

Where do tasks come from?

Requires tasks constructed from labeled data



Requires demos for many previous tasks



Requires labeled data from other regions

Rußwurm et al. Meta-Learning for Few-Shot Land Cover Classification. 2020

What if we only have unlabeled data?

few-shot meta-learning from: unlabeled images unlabeled text

A general recipe for unsupervised meta-learning

Given unlabeled dataset(s) → Propose tasks → Run meta-learning

Goal of unsupervised meta-learning methods:
Automatically construct tasks from unlabeled data

Question: What do you want the task set to look like?

1. diverse (more likely to cover test tasks)
2. structured (so that few-shot meta-learning is possible)

Next:

Task construction from unlabeled image data

Task construction from unlabeled text data

Can we meta-learn with only **unlabeled** images?

— — Task construction — —

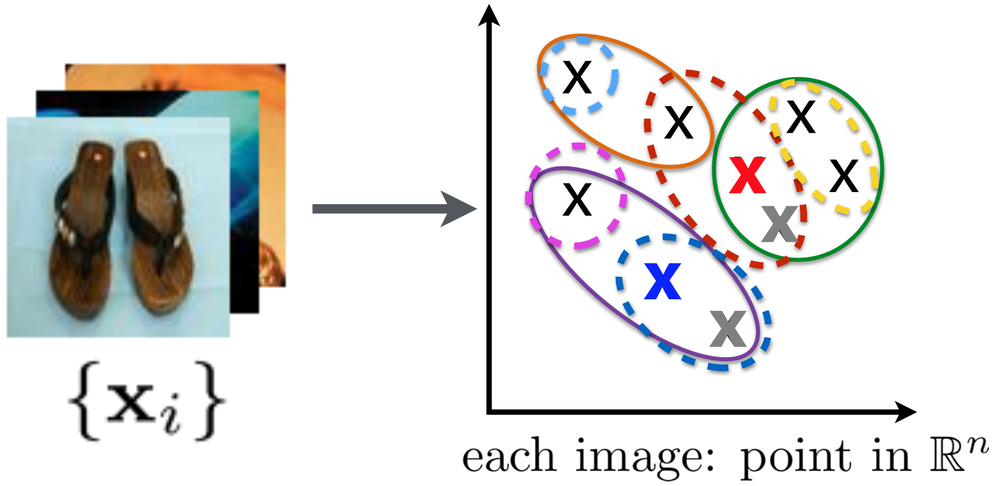
Unsupervised learning
(to get an embedding space)

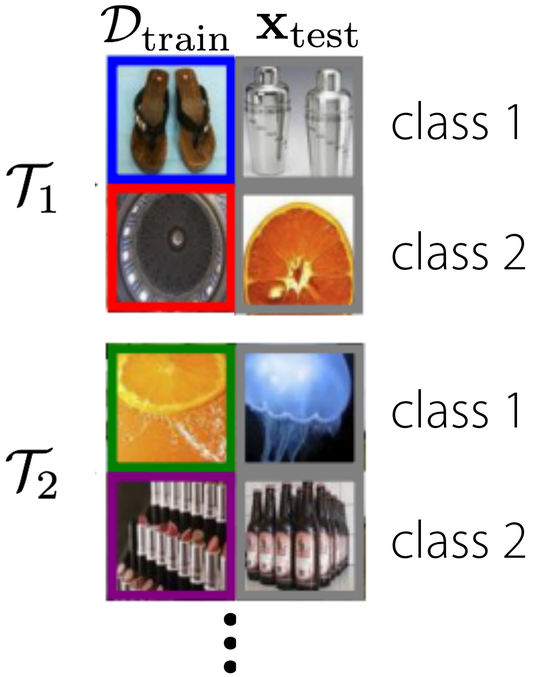
→

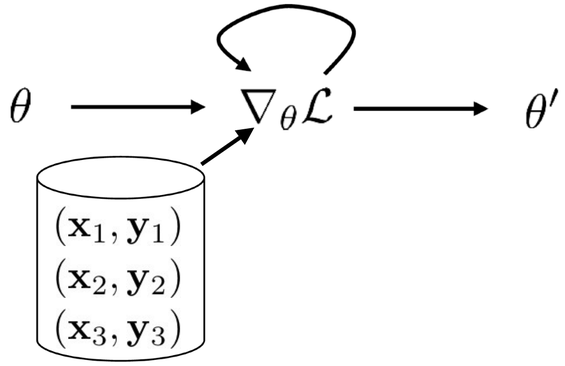
Propose cluster discrimination tasks

→

Run meta-learning







Result: representation suitable for learning downstream tasks

Hsu, Levine, Finn. **Unsupervised Learning via Meta-Learning**. ICLR'19

Can we meta-learn with only **unlabeled** images?

Unsupervised learning
(to get an embedding space)

A few options:
BiGAN — Donahue et al. '17
DeepCluster — Caron et al. '18

→

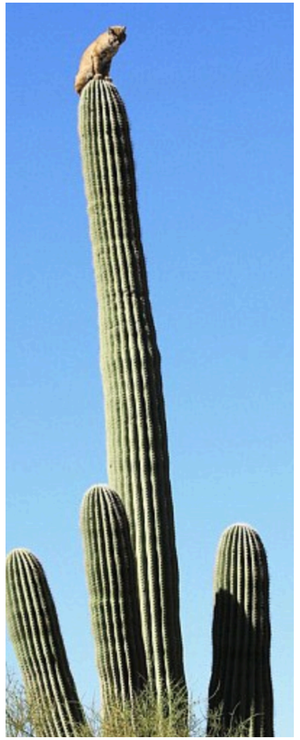
Propose cluster discrimination tasks

Clustering to Automatically Construct Tasks for Unsupervised Meta-Learning (CACTUs)

→

Run meta-learning

MAML — Finn et al. '17
ProtoNets — Snell et al. '17



CACTUs MAML

miniImageNet 5-way 5-shot

method	accuracy
MAML with labels	62.13%
BiGAN kNN	31.10%
BiGAN logistic	33.91%
BiGAN MLP + dropout	29.06%
BiGAN cluster matching	29.49%
BiGAN CACTUs MAML	51.28%
DeepCluster CACTUs MAML	**53.97%**

Same story for:

- **4 different embedding methods**
- **4 datasets** (Omniglot, CelebA, miniImageNet, MNIST)
- **2 meta-learning methods** (*)
- Test tasks with **larger datasets**

*ProtoNets underperforms in some cases.

Hsu, Levine, Finn. **Unsupervised Learning via Meta-Learning**. ICLR '19

Can we use **domain knowledge** when constructing tasks?

e.g. **image's label** often **won't change** when you:

- drop out some pixels
- translate the image
- reflect the image

Task construction:

For each task $\mathcal{T}_i$:

i. Randomly sample N images & assign labels $1,\ldots,N$

—> Store in $\mathcal{D}_i^{\text{tr}}$

ii. For each datapoint in $\mathcal{D}_i^{\text{tr}}$, augment image using domain knowledge

—> Store in $\mathcal{D}_i^{\text{ts}}$

Khodadadeh, Bölöni, Shah. **Unsupervised Meta-Learning for Few-Shot Image Classification**. NeurIPS '19

Can we use **domain knowledge** when constructing tasks?

For each task $\mathcal{T}_i$:

i. Randomly sample N images & assign labels $1,\ldots,N$ —> Store in $\mathcal{D}_i^{\text{tr}}$

ii. For each datapoint in $\mathcal{D}_i^{\text{tr}}$, augment image using domain knowledge —> Store in $\mathcal{D}_i^{\text{ts}}$

How to augment in practice?

Omniglot: translation & random pixel dropout **MiniImagenet:** AutoAugment* (translation, rotation, shear)

		Omniglot				Mini-Imagenet			
Algorithm (N, K)	**Clustering**	**(5,1)**	**(5,5)**	**(20,1)**	**(20,5)**	**(5,1)**	**(5,5)**	**(5,20)**	**(5,50)**
Training from scratch	N/A	52.50	74.78	24.91	47.62	27.59	38.48	51.53	59.63
linear classifier	ACAI / DC	61.08	81.82	43.20	66.33	29.44	39.79	56.19	65.28
MLP with dropout	ACAI / DC	51.95	77.20	30.65	58.62	29.03	39.67	52.71	60.95
cluster matching	ACAI / DC	54.94	71.09	32.19	45.93	22.20	23.50	24.97	26.87
CACTUs-MAML	ACAI / DC	68.84	87.78	48.09	73.36	39.90	**53.97**	**63.84**	**69.64**
CACTUs-ProtoNets	ACAI / DC	68.12	83.58	47.75	66.27	39.18	53.36	61.54	63.55
UMTRA (ours)	N/A	**83.80**	**95.43**	**74.25**	**92.12**	**39.93**	50.73	61.11	67.15
MAML (Supervised)	N/A	94.46	98.83	84.60	96.29	46.81	62.13	71.03	75.54
ProtoNets (Supervised)	N/A	98.35	99.58	95.31	98.81	46.56	62.29	70.05	72.04

- outstanding Omniglot performance
  (where we have good domain knowledge!)
- MiniImageNet: slightly underperforms CACTUs

Khodadadeh, Bölöni, Shah. **Unsupervised Meta-Learning for Few-Shot Image Classification**. NeurIPS '19

* Cubuk et al. 2018

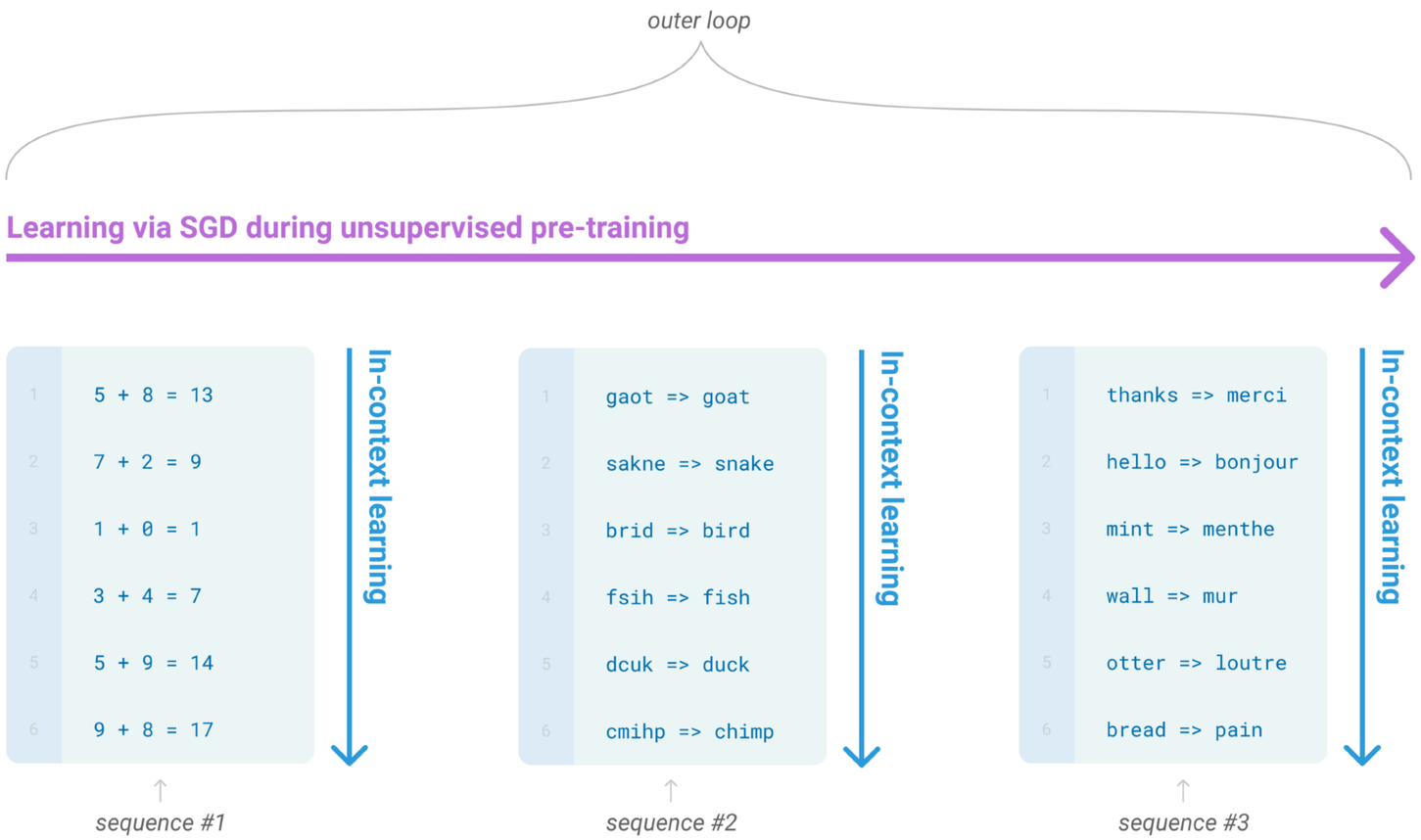
Can we meta-learn with only **unlabeled** text?

Option A: Formulate it as a language modeling problem.

Recall: GPT-3

$\mathscr{D}_i^{\text{tr}}$: sequence of characters

$\mathscr{D}_i^{\text{ts}}$: following sequence of characters



simple math problems | spelling correction | translating between languages

When might we not use this option?

- harder to combine w/ **optimization-based meta-learning**
- harder to apply to **classification** tasks (e.g. sentiment, political bias, etc)

Brown, Mann, Ryder, Subbiah et al. **Language Models are Few-Shot Learners**. arXiv '20

Can we meta-learn with only **unlabeled** text?

Option B: Construct tasks by masking out words

Task: Classify the masked word.

For each task $\mathcal{T}_i$:

i. Sample subset of N unique words & assign unique ID.

{Democratic, Capital} 1 2

ii. Sample $K + Q$ sentences with that word, *masking the word out*

iii. Construct $\mathcal{D}_i^{\text{tr}}$ and $\mathcal{D}_i^{\text{ts}}$ with masked sentences & corresponding word IDs

$\mathcal{D}_i^{\text{tr}}$

Support set

Sentence	Class
A member of the [m] Party, he was the first African American to be elected to the presidency.	1
The [m] Party is one of the two major contemporary political parties in the United States, along with its rival, the Republican Party.	1
Honolulu is the [m] and largest city of the U.S. state of Hawaii.	2
Washington, D.C., formally the District of Columbia and commonly referred to as Washington or D.C., is the [m] of the United States.	2

$\mathcal{D}_i^{\text{ts}}$

Query: New Delhi is an urban district of Delhi which serves as the [m] of India

Correct Prediction: 2

Bansal, Jha, Munkhdalai, McCallum. **Self-Supervised Meta-Learning for Few-Shot Natural Language Classification Tasks**. EMNLP '20

entirely unsupervised pre-training	supervised or semi-supervised pre-training

Task	N	k	BERT	**SMLMT**	MT-BERT$_{softmax}$	MT-BERT	LEOPARD	**Hybrid-SMLMT**
CoNLL	4	4	50.44 ± 08.57	46.81 ± 4.77	52.28 ± 4.06	55.63 ± 4.99	54.16 ± 6.32	**57.60** ± 7.11
		8	50.06 ± 11.30	61.72 ± 3.11	65.34 ± 7.12	58.32 ± 3.77	67.38 ± 4.33	**70.20** ± 3.00
		16	74.47 ± 03.10	75.82 ± 4.04	71.67 ± 3.03	71.29 ± 3.30	76.37 ± 3.08	**80.61** ± 2.77
		32	83.27 ± 02.14	84.01 ± 1.73	73.09 ± 2.42	79.94 ± 2.45	83.61 ± 2.40	**85.51** ± 1.73
MITR	8	4	49.37 ± 4.28	46.23 ± 3,90	45.52 ± 5.90	50.49 ± 4.40	49.84 ± 3.31	**52.29** ± 4.32
		8	49.38 ± 7.76	61.15 ± 1.91	58.19 ± 2.65	58.01 ± 3.54	62.99 ± 3.28	**65.21** ± 2.32
		16	69.24 ± 3.68	69.22 ± 2.78	66.09 ± 2.24	66.16 ± 3.46	70.44 ± 2.89	**73.37** ± 1.88
		32	78.81 ± 1.95	78.82 ± 1.30	69.35 ± 0.98	76.39 ± 1.17	78.37 ± 1.97	**79.96** ± 1.48
Airline	3	4	42.76 ± 13.50	42.83 ± 6.12	43.73 ± 7.86	46.29 ± 12.26	54.95 ± 11.81	**56.46** ± 10.67
		8	38.00 ± 17.06	51.48 ± 7.35	52.39 ± 3.97	49.81 ± 10.86	61.44 ± 03.90	**63.05** ± 8.25
		16	58.01 ± 08.23	58.42 ± 3.44	58.79 ± 2.97	57.25 ± 09.90	62.15 ± 05.56	**69.33** ± 2.24
		32	63.70 ± 4.40	65.33 ± 3.83	61.06 ± 3.89	62.49 ± 4.48	67.44 ± 01.22	**71.21** ± 3.28
Disaster	2	4	55.73 ± 10.29	**62.26** ± 9.16	52.87 ± 6.16	50.61 ± 8.33	51.45 ± 4.25	55.26 ± 8.32
		8	56.31 ± 09.57	**67.89** ± 6.83	56.08 ± 7.48	54.93 ± 7.88	55.96 ± 3.58	63.62 ± 6.84
		16	64.52 ± 08.93	**72.86** ± 1.70	65.83 ± 4.19	60.70 ± 6.05	61.32 ± 2.83	70.56 ± 2.23
		32	73.60 ± 01.78	**73.69** ± 2.32	67.13 ± 3.11	72.52 ± 2.28	63.77 ± 2.34	71.80 ± 1.85
Emotion	13	4	09.20 ± 3.22	09.84 ± 1.09	09.41 ± 2.10	09.84 ± 2.14	11.71 ± 2.16	**11.90** ± 1.74
		8	08.21 ± 2.12	11.02 ± 1.02	11.61 ± 2.34	11.21 ± 2.11	12.90 ± 1.63	**13.26** ± 1.01
		16	13.43 ± 2.51	12.05 ± 1.18	13.82 ± 2.02	12.75 ± 2.04	13.38 ± 2.20	**15.17** ± 0.89
		32	16.66 ± 1.24	14.28 ± 1.11	13.81 ± 1.62	**16.88** ± 1.80	14.81 ± 2.01	16.08 ± 1.16
Political Bias	2	4	54.57 ± 5.02	57.72 ± 5.72	54.32 ± 3.90	54.66 ± 3.74	60.49 ± 6.66	**61.17** ± 4.91
		8	56.15 ± 3.75	63.02 ± 4.62	57.36 ± 4.32	54.79 ± 4.19	61.74 ± 6.73	**64.10** ± 4.03
		16	60.96 ± 4.25	**66.35** ± 2.84	59.24 ± 4.25	60.30 ± 3.26	65.08 ± 2.14	66.11 ± 2.04
		32	65.04 ± 2.32	**67.73** ± 2.27	62.68 ± 3.21	64.99 ± 3.05	64.67 ± 3.41	67.30 ± 1.53

BERT - standard self-supervised learning + fine-tuning

SMLMT - proposed unsupervised meta-learning

MT-BERT - multi-task learning + fine-tuning (on supervised tasks)

LEOPARD - optimization-based meta-learner (only on supervised tasks)

Hybrid-SMLMT - meta-learning on proposed tasks + supervised tasks

More results & analysis in the paper!

Bansal, Jha, Munkhdalai, McCallum. **Self-Supervised Meta-Learning for Few-Shot Natural Language Classification Tasks**. EMNLP '20

Plan for Today

Brief Recap of Meta-Learning & Task Construction

Memorization in Meta-Learning
- When it arises
- A potential solutions

Meta-Learning without Tasks
- Unsupervised Meta-Learning
- **Meta-Learning from Unsegmented Task Stream** (time permitting)

What if we have a time series of labeled data?

- predict energy demand
- dynamics of a robot, car
- transportation usage
- stock market
- video analytics
- RL agent

unsegmented
yet, exhibits **temporal structure**

Can we **segment time series** into tasks & **meta-learn** across tasks?

How to segment?

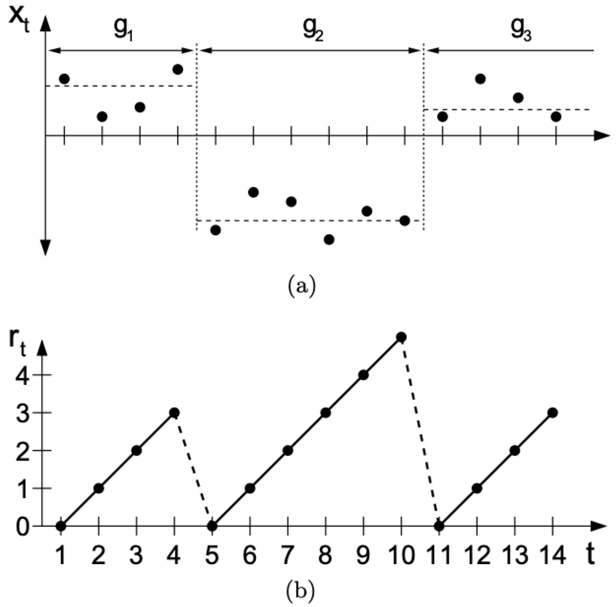
Bayesian online change point detection (BOCPD)

Adams & Mackay '17

Problem: assume task will switch with some probability, at each time t

Maintain **belief over task duration** (run length), **posterior** for each duration

Recursively update belief using **model performance**



BOCPD is differentiable! —> backprop through update belief update to meta-train model

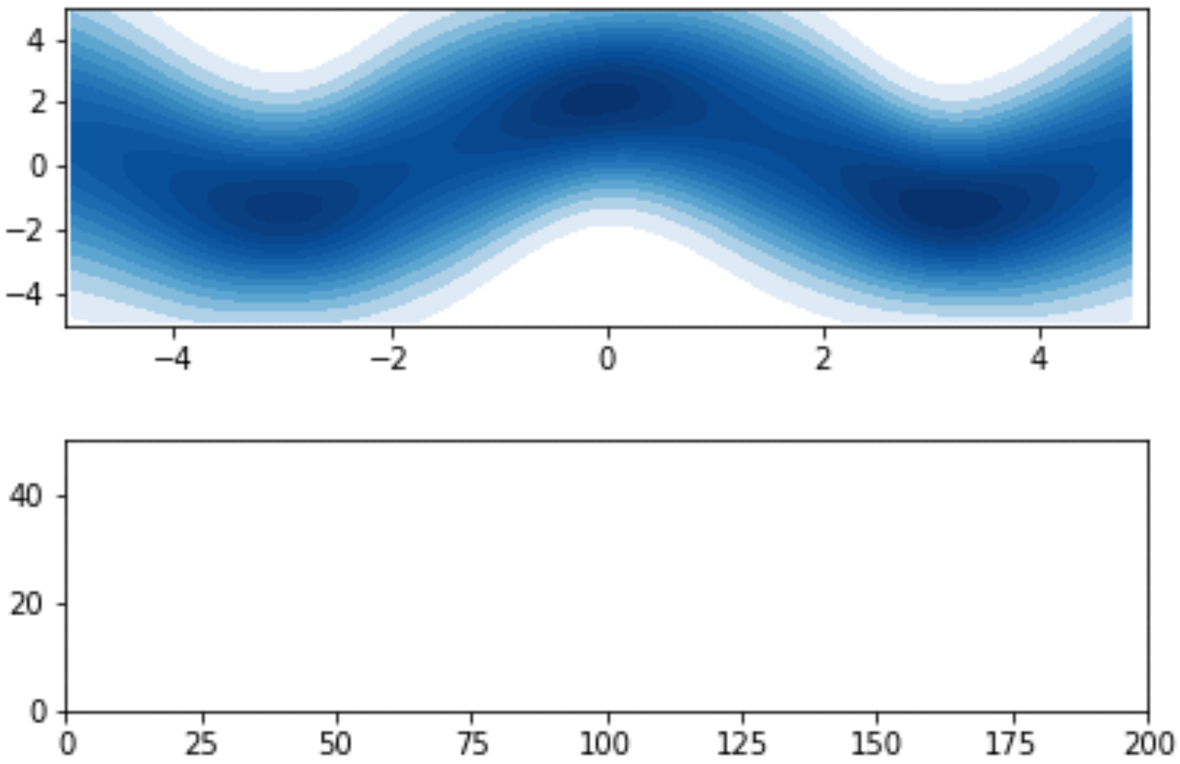
Harrison, Sharma, Finn, Pavone. **Continuous Meta-Learning without Tasks**. NeurIPS '20

Meta-Learning with Online Changepoint Analysis (MOCA)

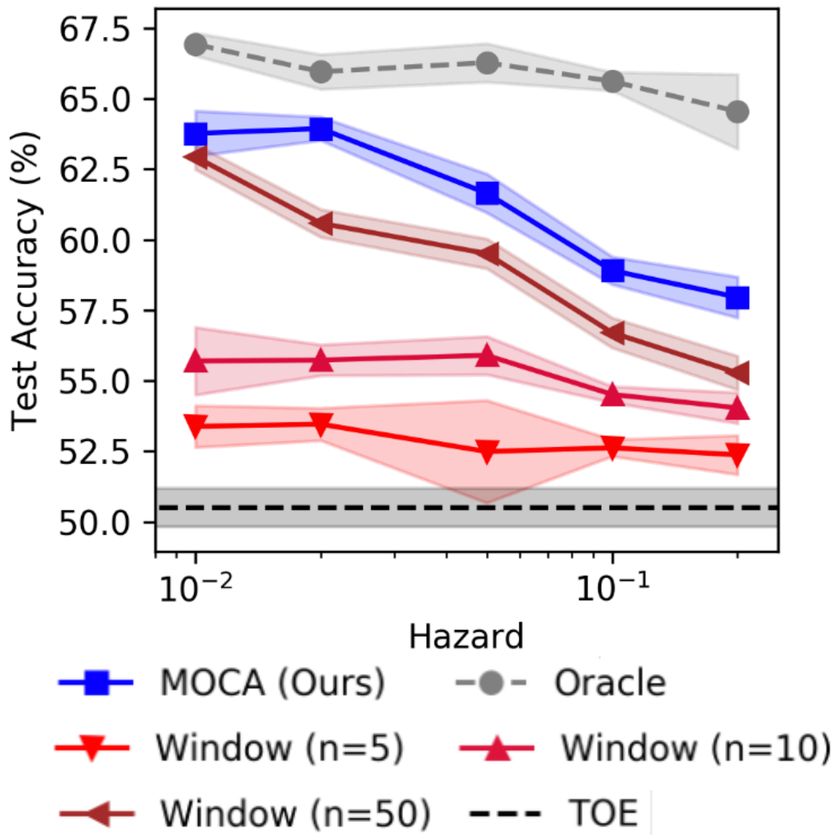
Meta-training phase: given unsegmented time-series of offline data

Meta-test phase: streaming online learning & prediction

Sinusoid regression with discrete shifts



Streaming variant of MiniImagenet.



Harrison, Sharma, Finn, Pavone. **Continuous Meta-Learning without Tasks**. NeurIPS '20

Plan for Today

Brief Recap of Meta-Learning & Task Construction

Memorization in Meta-Learning
- When it arises
- A potential solutions

Meta-Learning without Tasks Provided
- Unsupervised Meta-Learning
- Meta-Learning from Unsegmented Task Stream (time permitting)

🚨 **Disclaimer** 🚨**:** These topics are at the bleeding edge of research.

Goals for by the end of lecture:
- Understand when & how memorization in meta-learning may occur
- Understand techniques for constructing tasks automatically

Reminders

Today:

Homework 1 due, Homework 2 out

Project survey due (**non-binding**, for assigning **mentor/PoC**)

Tomorrow at 4 pm:

Tutorial on variational inference