Optimization-Based Meta-Learning

CS 330

Course Reminders

Project group form due **tonight**.

(for assigning project mentors)

Homework 1 due **Monday**

Homework 2 out **today**

Tutorial session **tomorrow 4:30-5:20 pm** on MAML.

Guest lectures!

James Harrison

Google DeepMind

Learned optimizers

Jason Wei

OpenAI

In-context learning

Plan for Today

Recap
- Meta-learning problem & black-box meta-learning

Optimization Meta-Learning } **Part of Homework 2!**
- Overall approach
- Compare: optimization-based vs. black-box
- Challenges & solutions
- Case study of land cover classification (time-permitting)

Goals for by the end of lecture:
- Basics of optimization-based meta-learning techniques (& how to implement)
- Trade-offs between black-box and optimization-based meta-learning

Problem Settings Recap

Multi-Task Learning

Solve multiple tasks $\mathcal{T}_1, \cdots, \mathcal{T}_T$ at once.

$$\min_{\theta} \sum_{i=1}^{T} \mathcal{L}_i(\theta, \mathcal{D}_i)$$

Transfer Learning

Solve target task $\mathcal{T}_b$ after solving source task $\mathcal{T}_a$ by *transferring* knowledge learned from $\mathcal{T}_a$

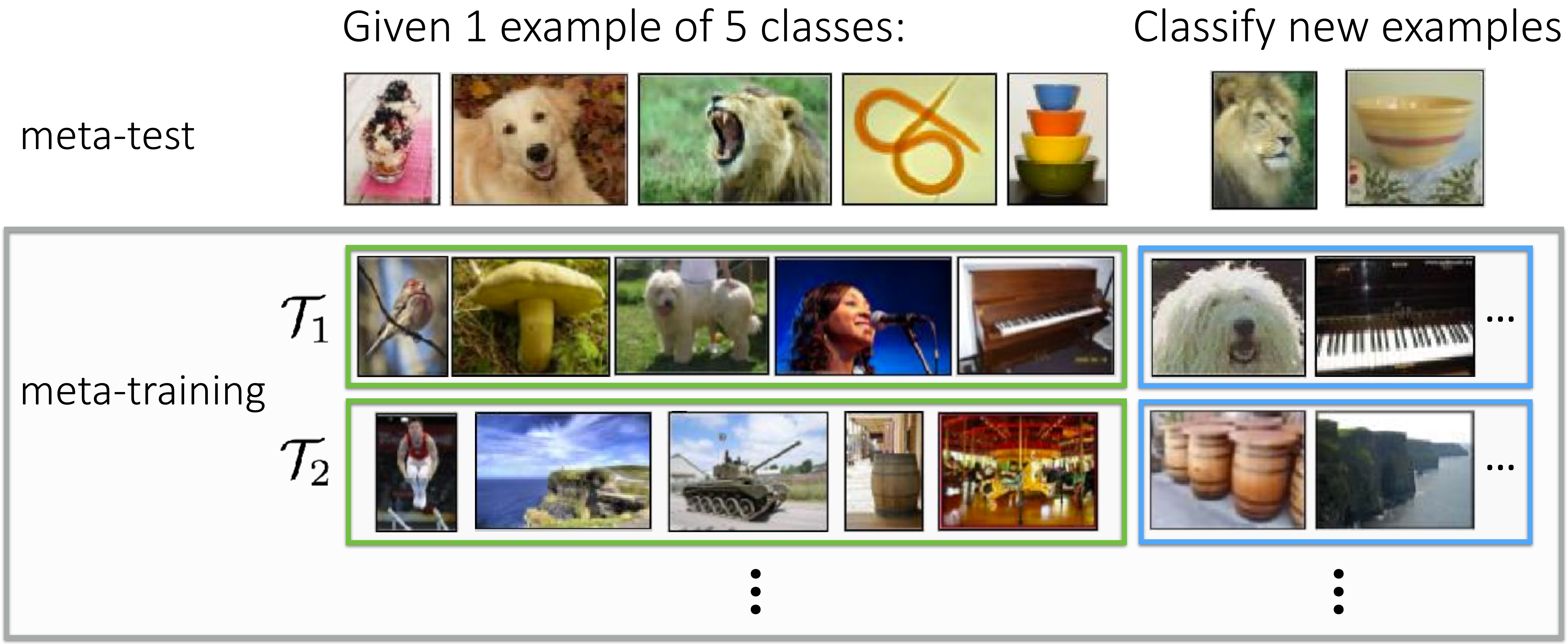
Meta-Learning Problem

Transfer Learning with Many Source Tasks

Given data from $\mathcal{T}_1, \ldots, \mathcal{T}_n$, solve new task $\mathcal{T}_{\text{test}}$ more quickly / proficiently / stably

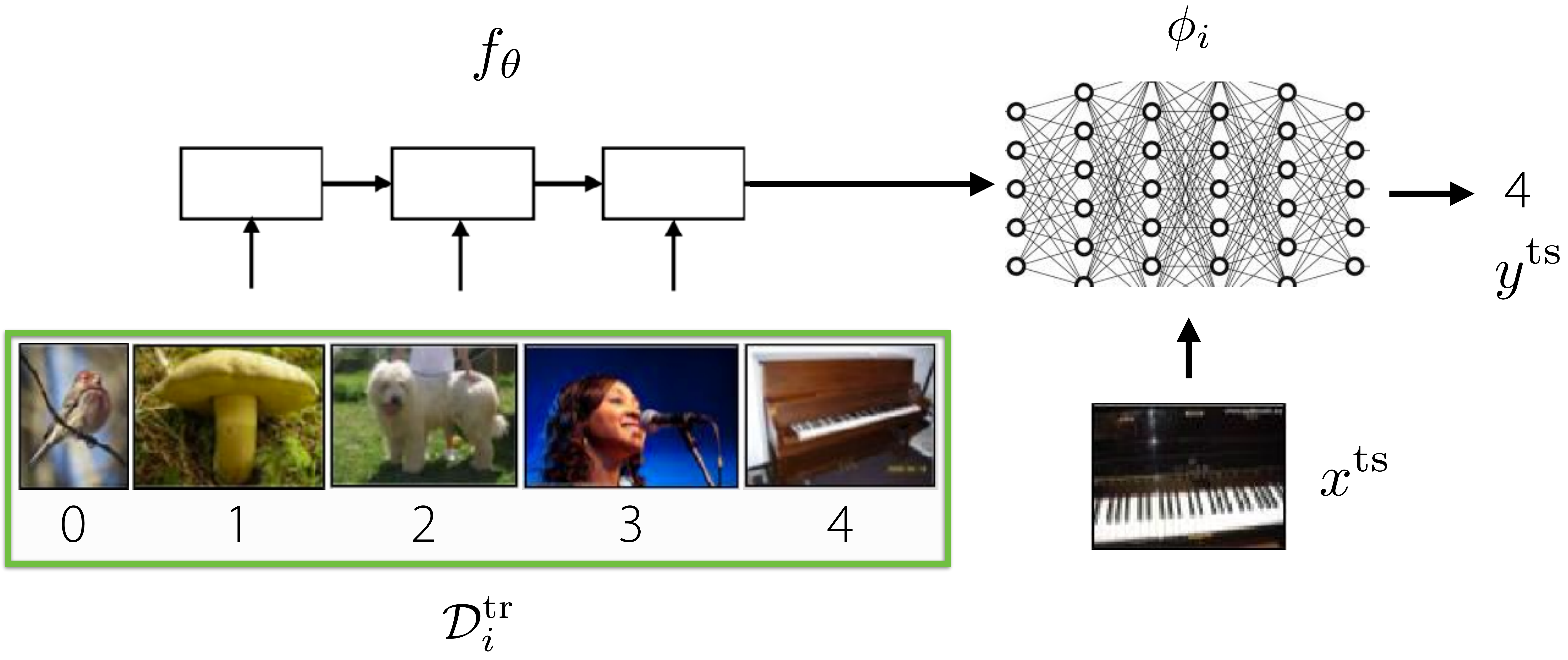
Example Meta-Learning Problem

5-way, **1**-shot image classification (MiniImagenet)



Can replace image classification with: **regression**, **language generation**, **skill learning**, **any ML problem**

Black-Box Adaptation



general form:

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

+ **expressive**

- **challenging optimization** problem

Case Study: GPT-3

Language Models are Few-Shot Learners

Tom B. Brown[*] Benjamin Mann[*] Nick Ryder[*] Melanie Subbiah[*]

Jared Kaplan[†] Prafulla Dhariwal Arvind Neelakantan Pranav Shyam Girish Sastry

Amanda Askell Sandhini Agarwal Ariel Herbert-Voss Gretchen Krueger Tom Henighan

Rewon Child Aditya Ramesh Daniel M. Ziegler Jeffrey Wu Clemens Winter

Christopher Hesse Mark Chen Eric Sigler Mateusz Litwin Scott Gray

Benjamin Chess Jack Clark Christopher Berner

Sam McCandlish Alec Radford Ilya Sutskever Dario Amodei

OpenAI

May 2020

"emergent" few-shot learning

What is GPT-3?

a language model

black-box meta-learner *trained on language generation tasks*

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

[meta-training] dataset: crawled data from the internet, English-language Wikipedia, two books corpora

architecture: giant "Transformer" network 175 billion parameters, 96 layers, 3.2M batch size

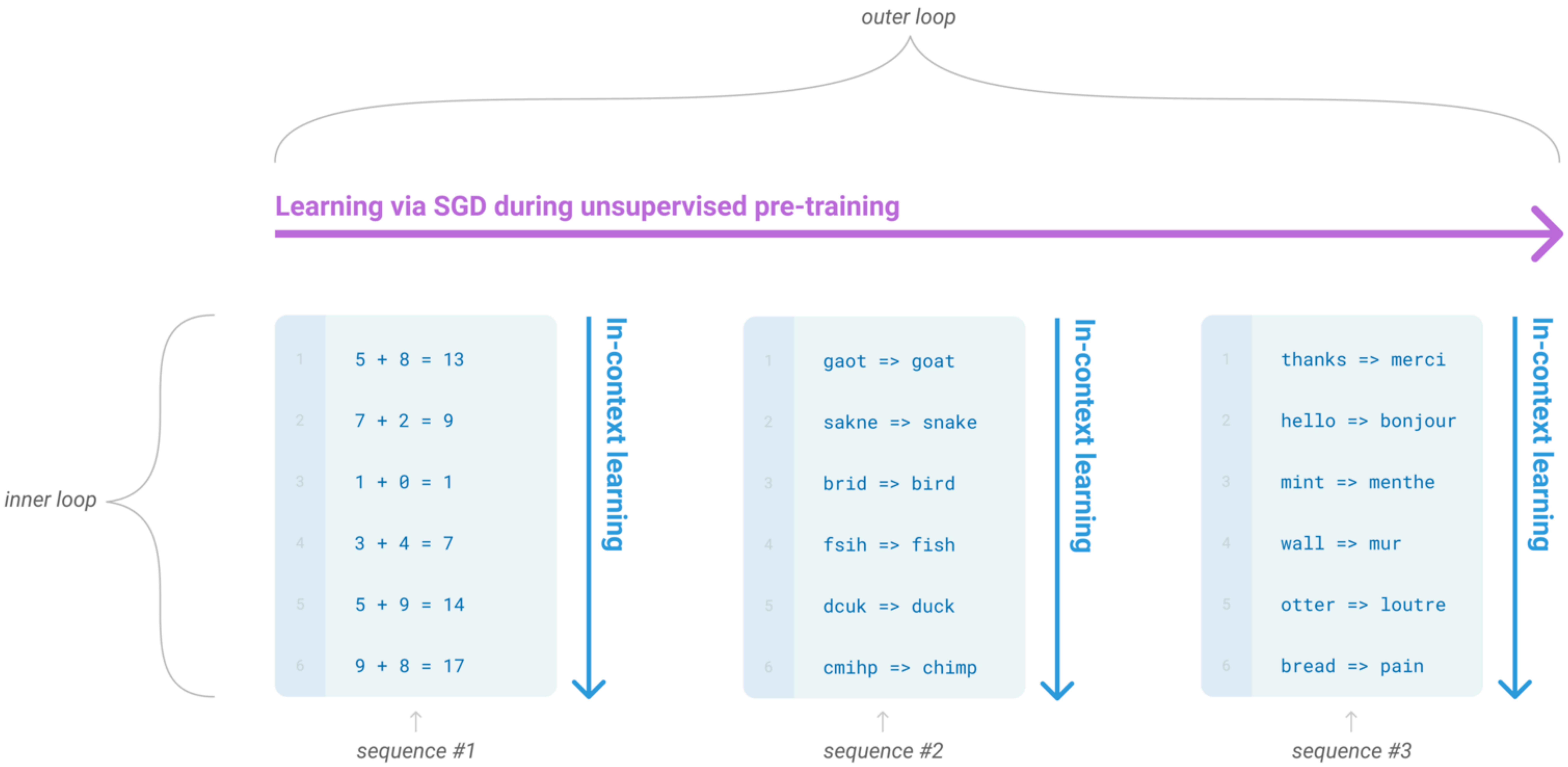
What do different tasks correspond to?

- spelling correction
- simple math problems
- translating between languages
- a variety of other tasks

How can those tasks all be solved by a single architecture?

How can those tasks all be solved by a single architecture? Put them all in the form of text!

Why is that a good idea? Very easy to get a lot of meta-training data.



simple math problems spelling correction translating between languages

Some Results

One-shot learning from dictionary definitions:

To "screeg" something is to swing a sword at it. An example of a sentence that uses the word screeg is:
We screeghed at each other for several minutes and then we went outside and ate ice cream.

Few-shot language editing:

Poor English input: I eated the purple berries.
Good English output: I ate the purple berries.
Poor English input: Thank you for picking me as your designer. I'd appreciate it.
Good English output: Thank you for choosing me as your designer. I appreciate it.
Poor English input: The mentioned changes have done. or I did the alteration that you requested. or I changed things you wanted and did the modifications.
Good English output: The requested changes have been made. or I made the alteration that you requested. or I changed things you wanted and made the modifications.
Poor English input: I'd be more than happy to work with you in another project.
Good English output: I'd be more than happy to work with you on another project.

Poor English input: Please provide me with a short brief of the design you're looking for and that'd be nice if you could share some examples or project you did before.
Good English output: Please provide me with a brief description of the design you're looking for and that would be nice if you could share some examples or projects you have done before.

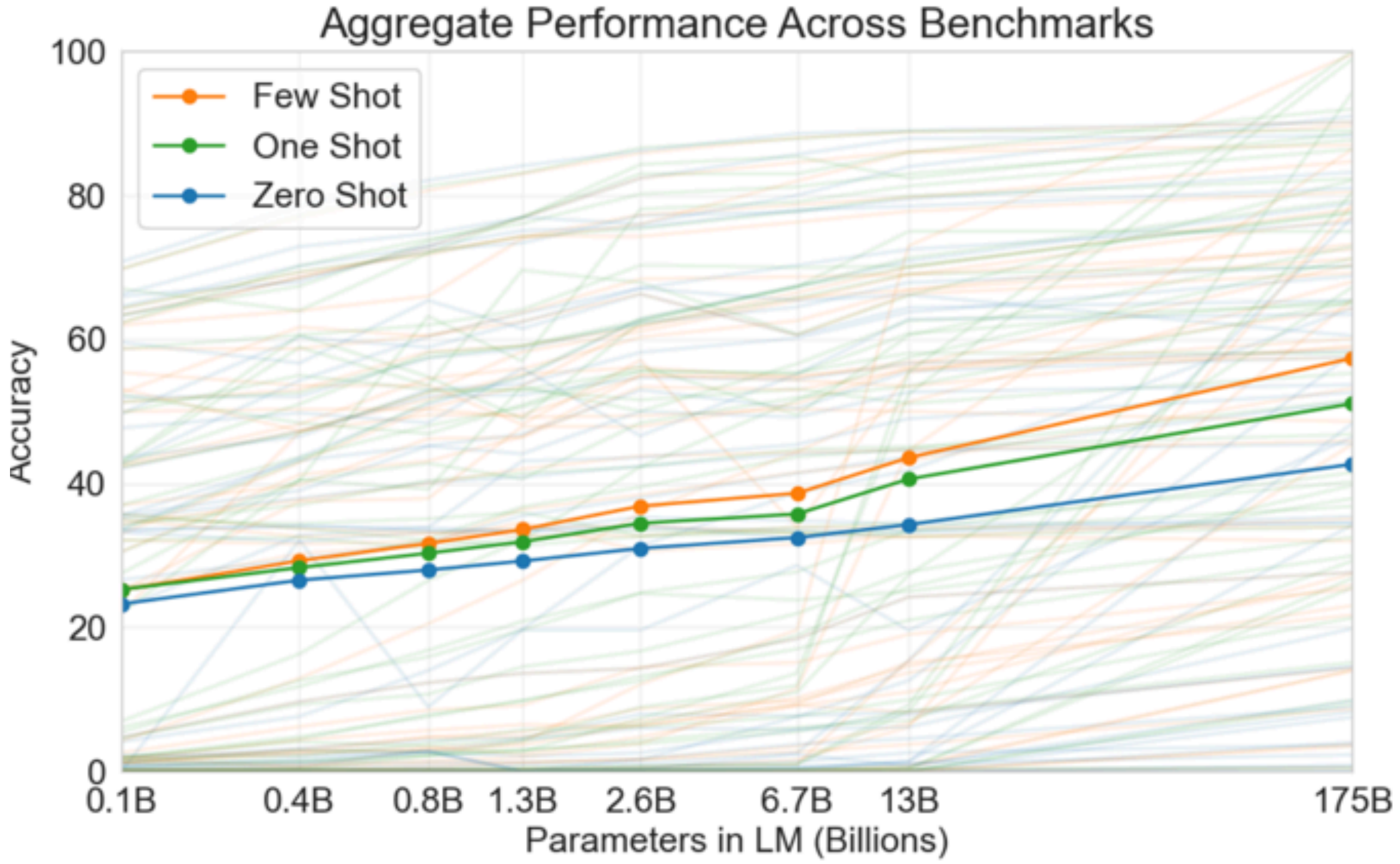
Non-few-shot learning tasks:

Title: United Methodists Agree to Historic Split
Subtitle: Those who oppose gay marriage will form their own denomination
Article: **After two days of intense debate, the United Methodist Church has agreed to a historic split - one that is expected to end in the creation of a new denomination, one that will be "theologically and socially conservative," according to The Washington Post. The majority of delegates attending the church's annual General Conference in May voted to strengthen a ban on the ordination of LGBTQ clergy and to write new rules that will "discipline" clergy who officiate at same-sex weddings. But those who opposed these measures have a new plan: They say they will form a separate denomination by 2020, calling their church the Christian Methodist**

General Notes & Takeaways

The results are extremely impressive.

The model is far from perfect.



The model fails in unintuitive ways.

```
Q: How many eyes does a giraffe have?
A: A giraffe has two eyes.

Q: How many eyes does my foot have?
A: Your foot has two eyes.

Q: How many eyes does a spider have?
A: A spider has eight eyes.

Q: How many eyes does the sun have?
A: The sun has one eye.
```

Source: https://lacker.io/ai/2020/07/06/giving-gpt-3-a-turing-test.html

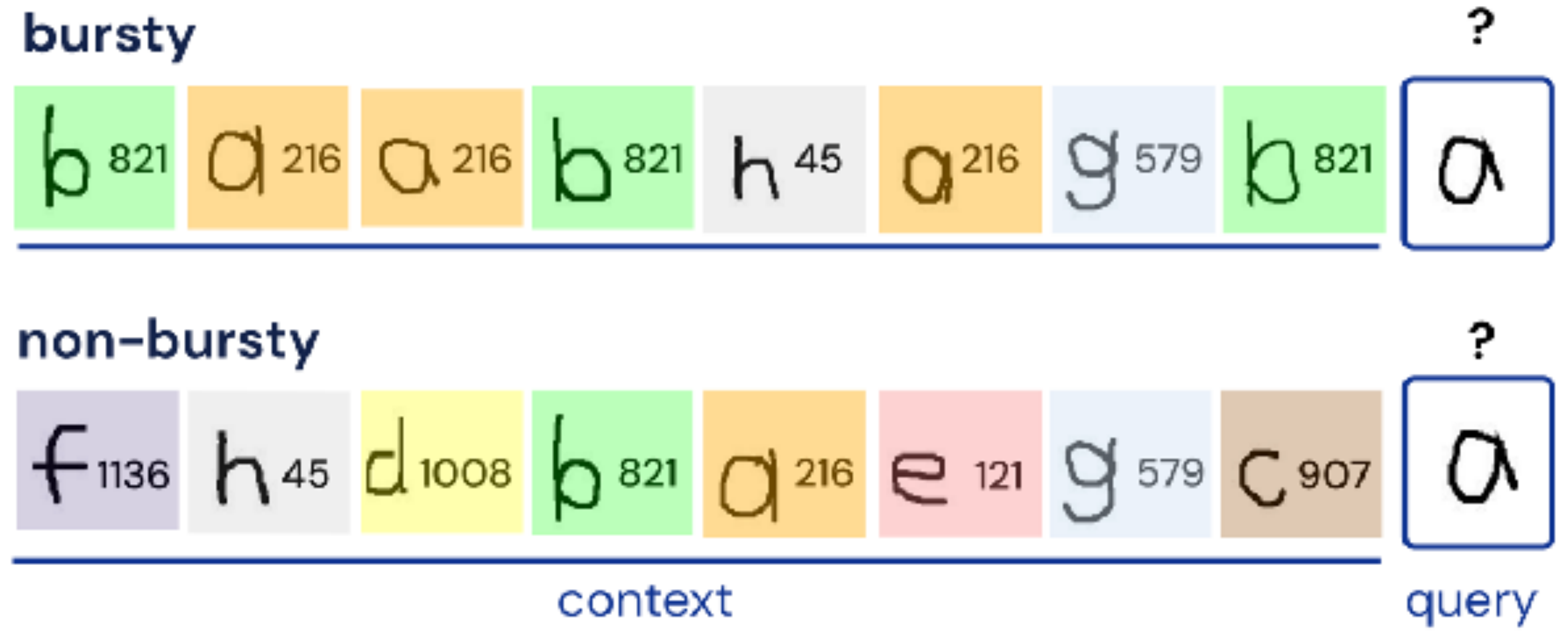
The choice of $\mathscr{D}_i^{tr}$ at test time is important. ("prompting")

Source: https://github.com/shreyashankar/gpt3-sandbox/blob/master/docs/priming.md

What is needed for few-shot learning to emerge?

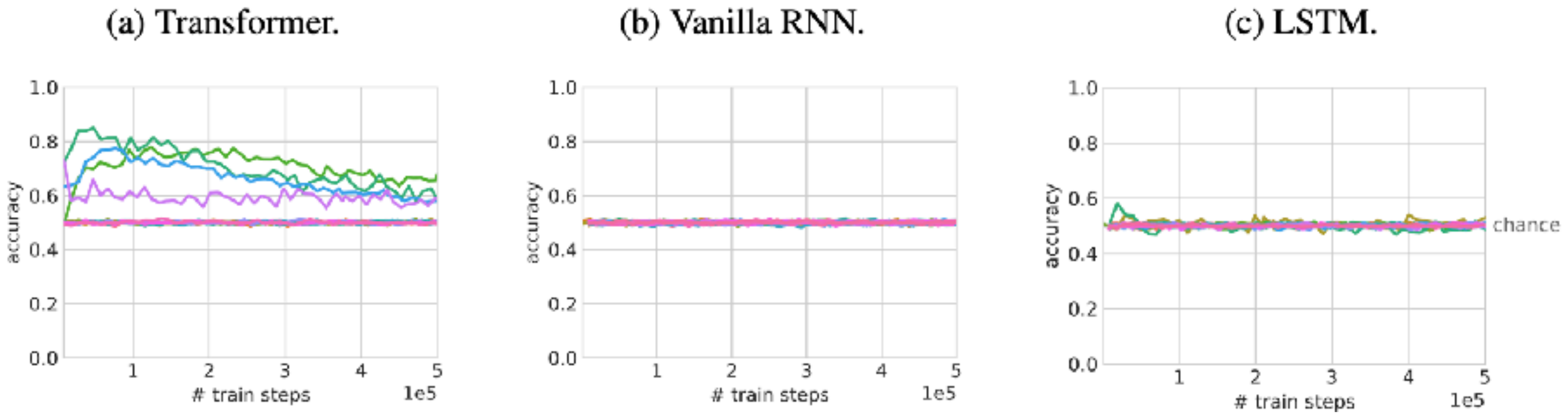
An active research topic!

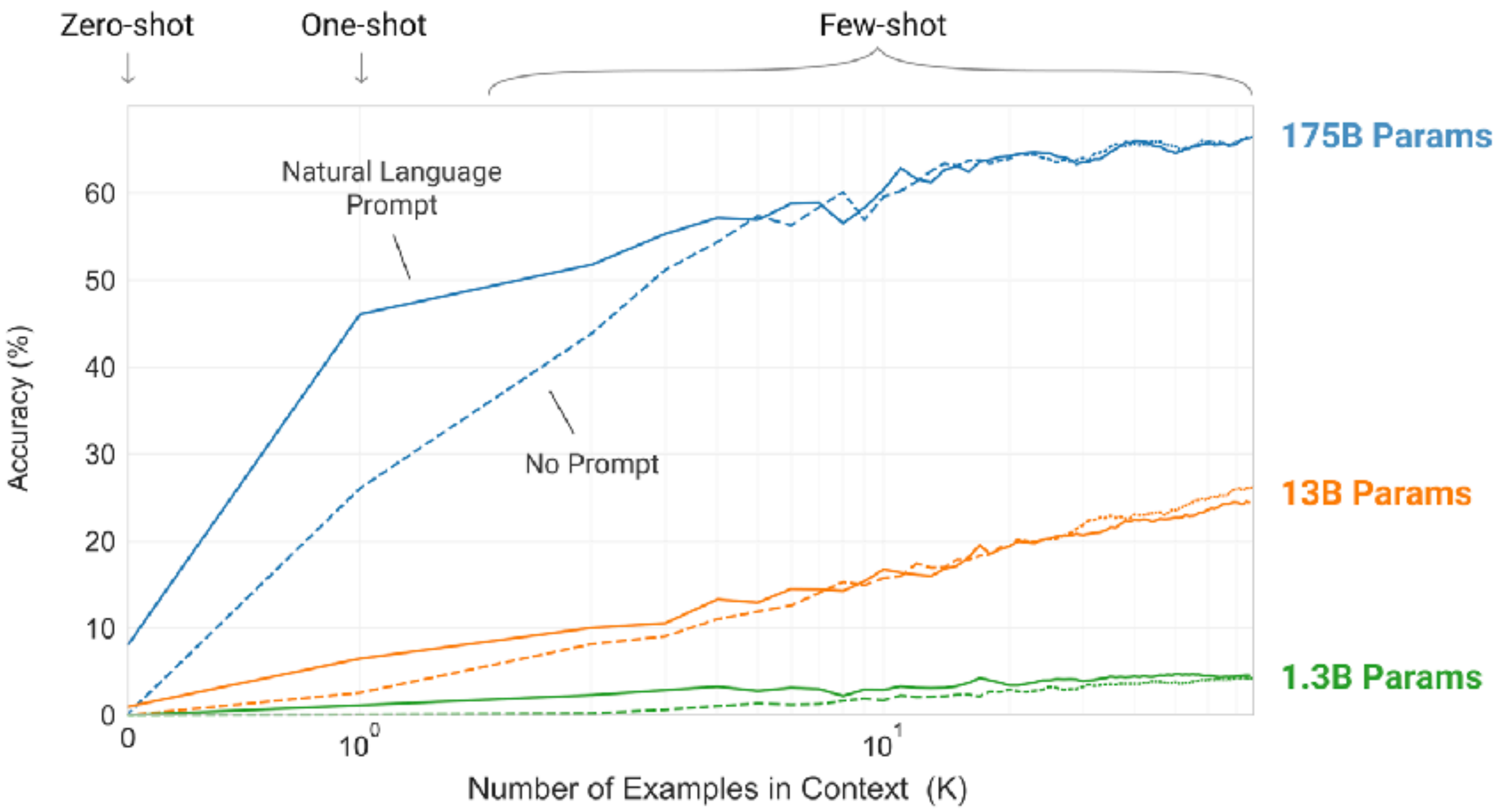
Data:
- temporal correlation
- dynamic meaning of words



Model:
- large capacity models

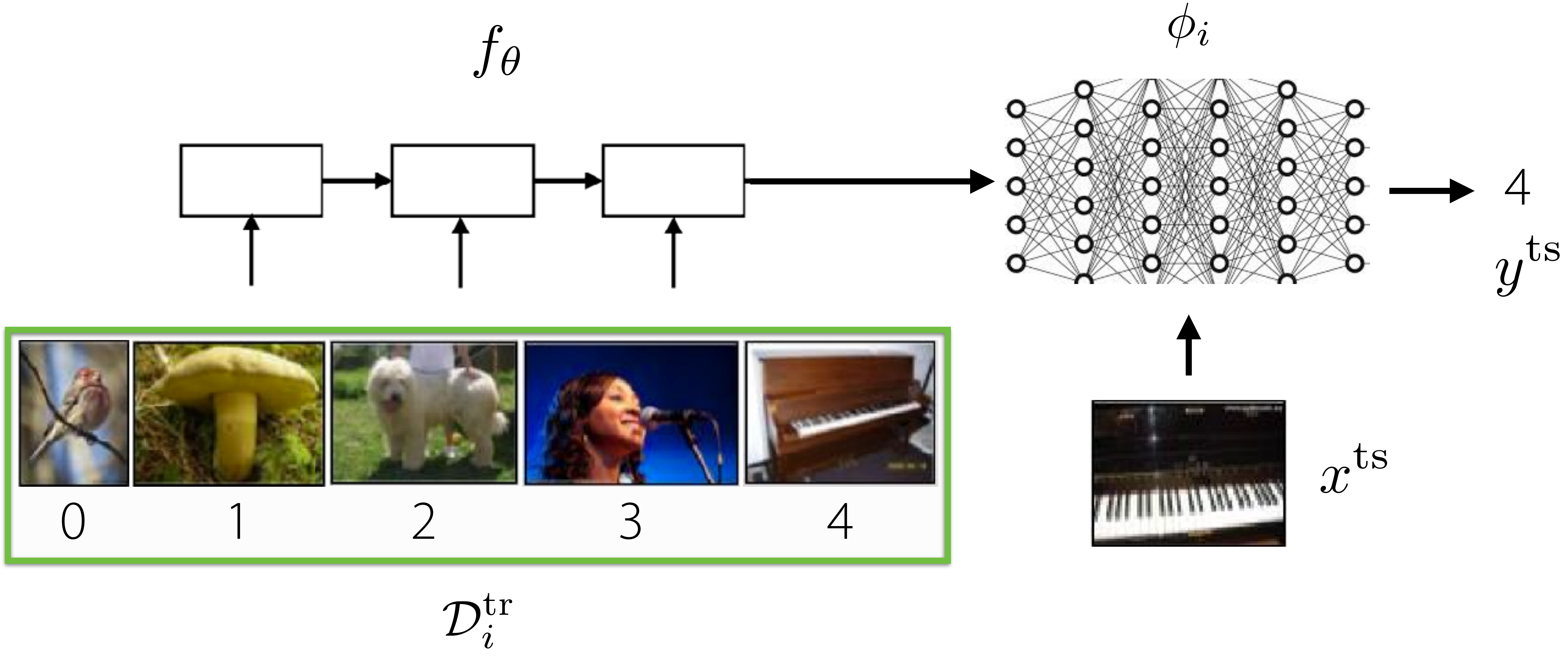
transformers > RNNs
large models > small models





Chan, Santoro, Lampinen, Wang, Singh, Richemond, McClelland, Hill. Data Distributional Properties Drive Emergent In-Context Learning in Transformers. '22
Brown*, Mann*, Ryder*, Subbiah* et al. Language Models are Few-Shot Learners. '20

Black-Box Adaptation



general form:

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

\+ **expressive**

\- **challenging optimization** problem

How else can we represent $\phi_i = f_\theta(\mathcal{D}_i^{\text{tr}})$?

What if we treat it as an **optimization** procedure?

Plan for Today

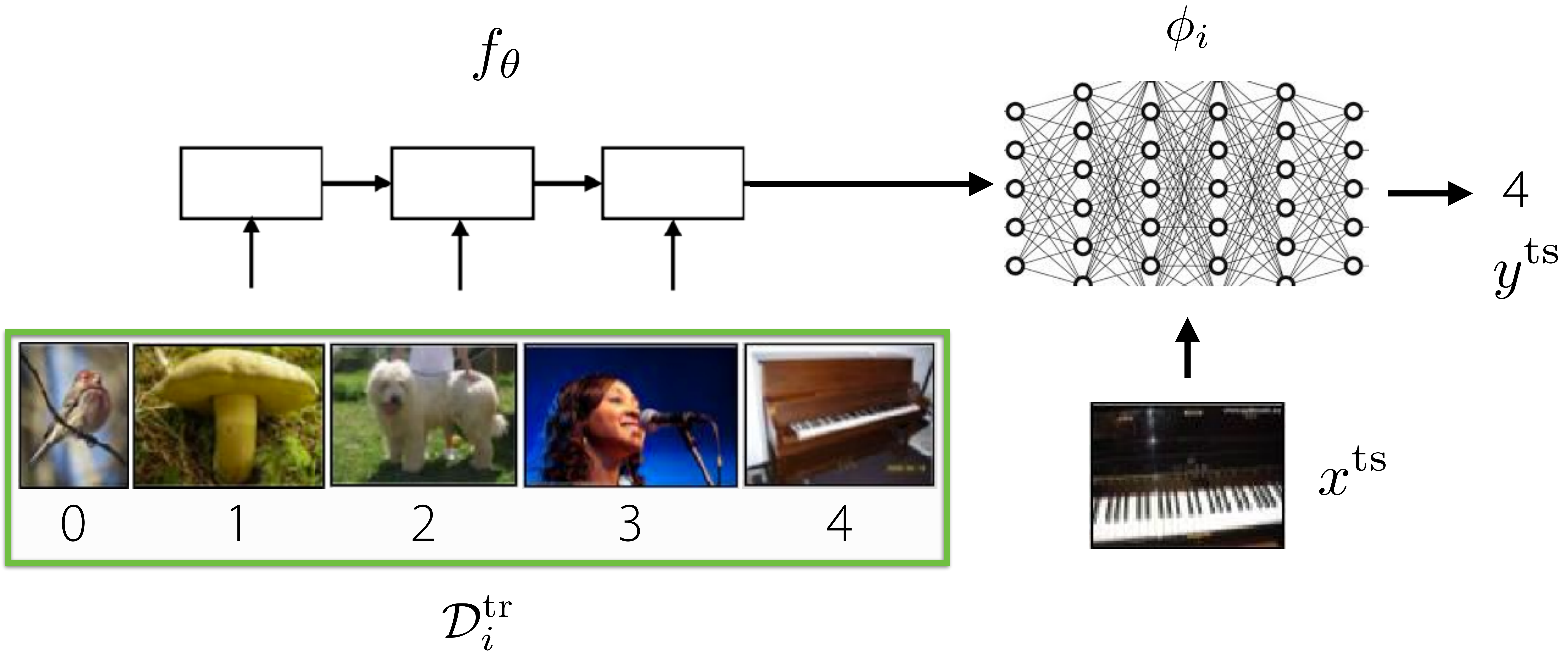
Recap

- Meta-learning problem & black-box meta-learning

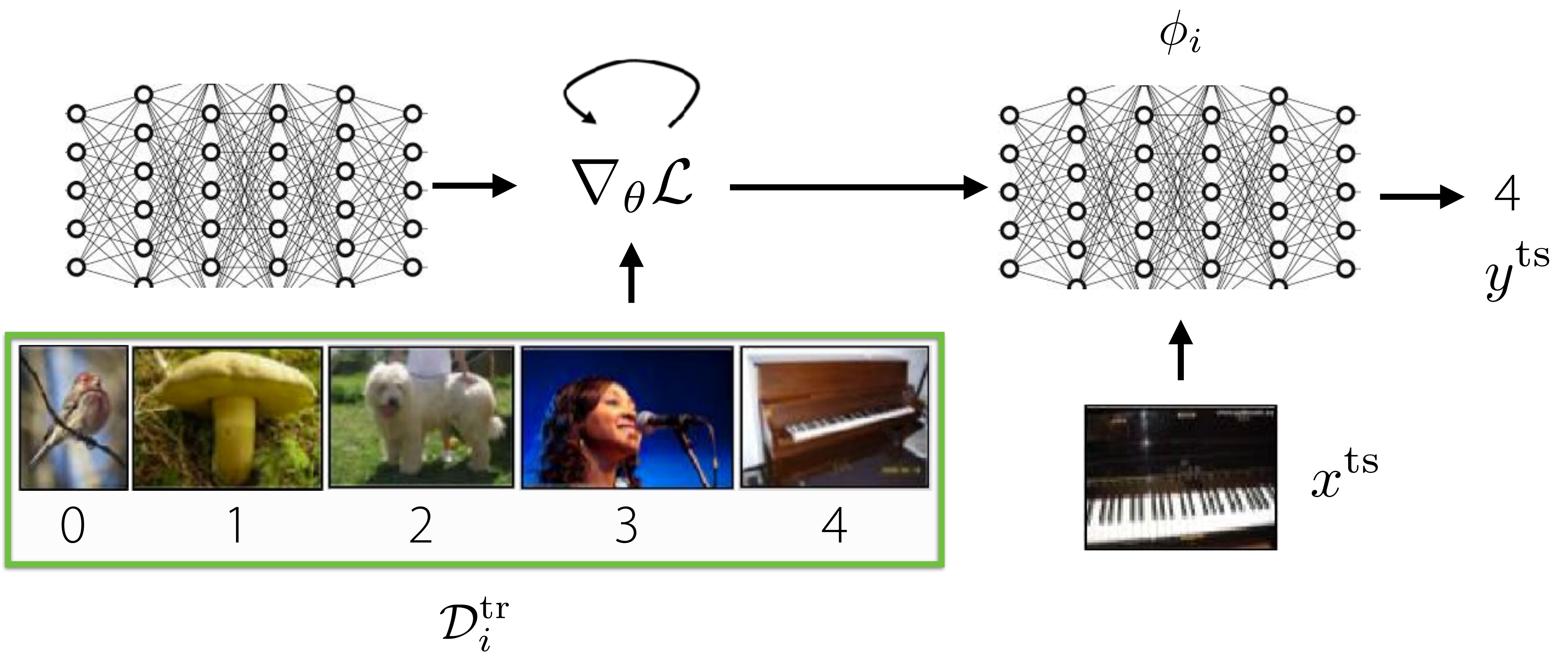
Optimization Meta-Learning } Part of Homework 2!

- Overall approach
- Compare: optimization-based vs. black-box
- Challenges & solutions
- Case study of land cover classification (time-permitting)

~~Black-Box Adaptation~~ Optimization-Based Adaptation



~~Black Box Adaptation~~ Optimization-Based Adaptation



Key idea: embed optimization inside the inner learning process

Why might this make sense?

Recall: Fine-tuning

pre-trained parameters

Fine-tuning $$\phi \leftarrow \theta - \alpha \nabla_\theta \mathcal{L}(\theta, \mathcal{D}^{\mathrm{tr}})$$

training data for new task

(typically for many gradient steps)

Universal Language Model Fine-Tuning for Text Classification. Howard, Ruder. '18

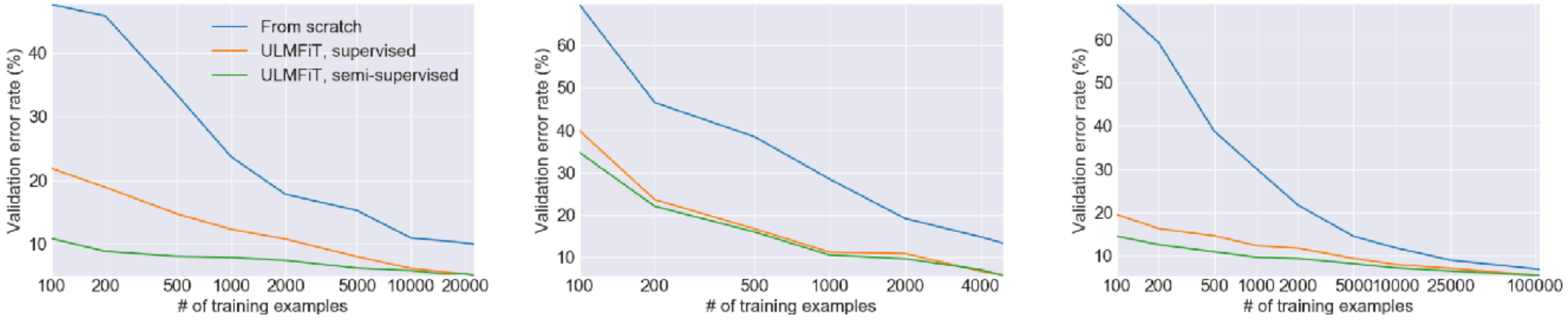


Figure 3: Validation error rates for supervised and semi-supervised ULMFiT vs. training from scratch with different numbers of training examples on IMDb, TREC-6, and AG (from left to right).

Fine-tuning less effective with **very small datasets**.

Optimization-Based Adaptation

pre-trained parameters

Fine-tuning
[test-time]

$$\phi \leftarrow \theta - \alpha \nabla_\theta \mathcal{L}(\theta, \mathcal{D}^{\text{tr}})$$

training data
for new task

Meta-learning $$\min_\theta \sum_{\text{task } i} \mathcal{L}(\theta - \alpha \nabla_\theta \mathcal{L}(\theta, \mathcal{D}_i^{\text{tr}}), \mathcal{D}_i^{\text{ts}})$$

Key idea: Over many tasks, learn parameter vector θ that transfers via fine-tuning

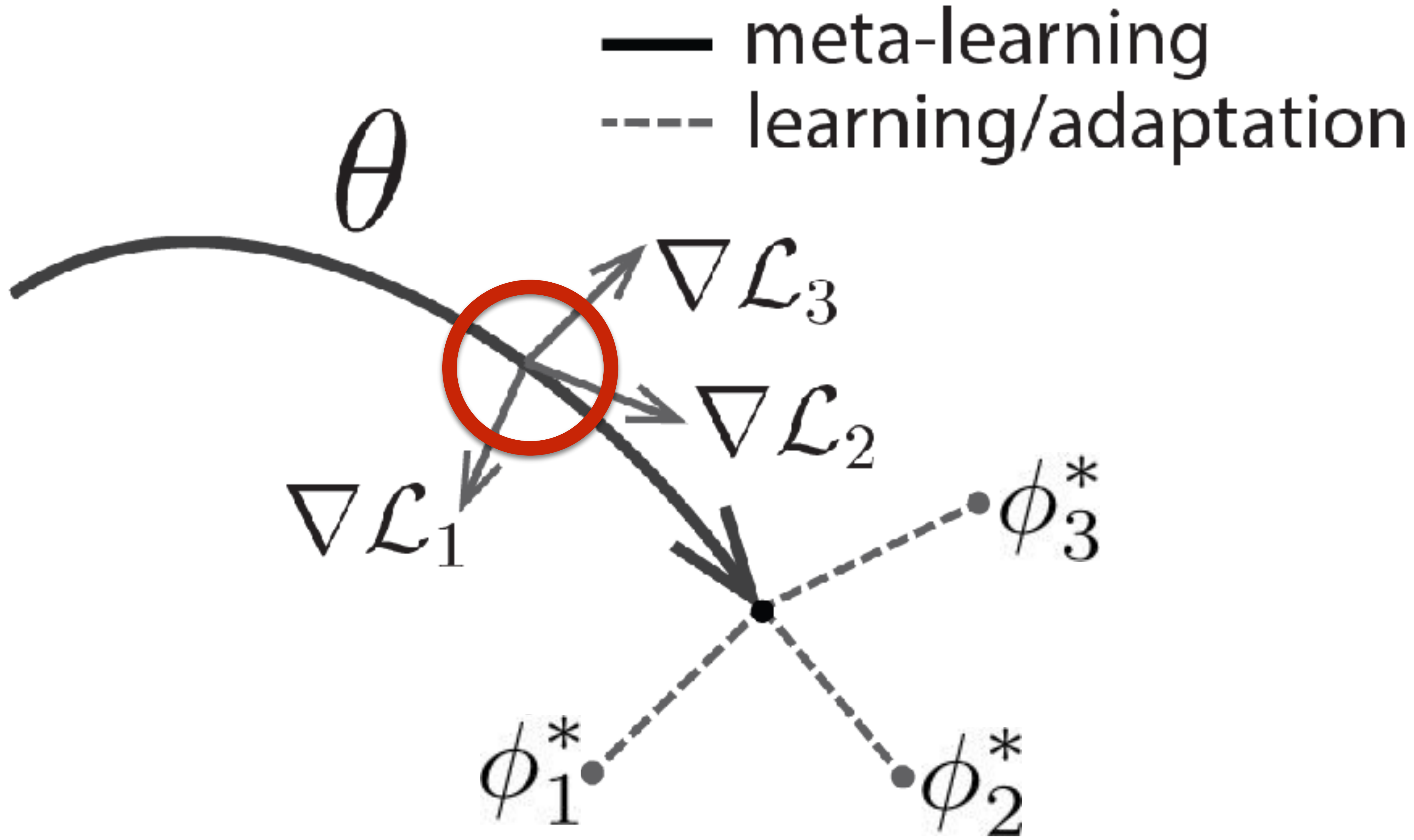
Finn, Abbeel, Levine. Model-Agnostic Meta-Learning. ICML 2017

Optimization-Based Adaptation

$$\min_{\theta} \sum_{\text{task } i} \mathcal{L}(\theta - \alpha \nabla_{\theta} \mathcal{L}(\theta, \mathcal{D}_i^{\text{tr}}), \mathcal{D}_i^{\text{ts}})$$

θ parameter vector being meta-learned

ϕ_i^* optimal parameter vector for task i



Model-Agnostic Meta-Learning

Finn, Abbeel, Levine. Model-Agnostic Meta-Learning. ICML 2017

Optimization-Based Adaptation

Key idea: Acquire ϕ_i through optimization.

General Algorithm:

~~Black box approach~~ Optimization-based approach

1. Sample task $\mathcal{T}_i$ *(or mini batch of tasks)*
2. Sample disjoint datasets $\mathcal{D}_i^{\text{tr}}, \mathcal{D}_i^{\text{test}}$ from $\mathcal{D}_i$
3. ~~Compute $\phi_i \leftarrow f_\theta(\mathcal{D}_i^{\text{tr}})$~~ Optimize $\phi_i \leftarrow \theta - \alpha \nabla_\theta \mathcal{L}(\theta, \mathcal{D}_i^{\text{tr}})$
4. Update θ using $\nabla_\theta \mathcal{L}(\phi_i, \mathcal{D}_i^{\text{test}})$

—> brings up **second-order** derivatives

Do we need to compute the full Hessian?

-> whiteboard

Do we get higher-order derivatives with more inner gradient steps?

$$\begin{aligned}
&\frac{d}{d\theta}\mathcal{L}(\phi_i, \mathcal{D}_i^{\text{ts}}) \\
&= \nabla_{\bar{\phi}}\mathcal{L}(\bar{\phi}, \mathcal{D}_i^{\text{ts}})|_{\bar{\phi}=\phi_i} \frac{d\phi_i}{d\theta} \\
&= \nabla_{\bar{\phi}}\mathcal{L}(\bar{\phi}, \mathcal{D}_i^{\text{ts}})|_{\bar{\phi}=\phi_i} \left(I - \alpha \frac{d^2}{d\theta^2}\mathcal{L}(\theta, \mathcal{D}_i^{\text{tr}})\right)
\end{aligned}$$

PyTorch

Deep learning libraries handle the math for you.

Optimization-Based Adaptation

Key idea: Acquire ϕ_i through optimization.

Meta-Test Time:

Optimization-based approach

1. Given task $\mathcal{T}_j$
2. Given training data $\mathcal{D}_j^{\mathrm{tr}}$
3. Fine-tune $\phi_j \leftarrow \theta - \alpha \nabla_\theta \mathcal{L}(\theta, \mathcal{D}_j^{\mathrm{tr}})$
4. Make predictions on new datapoints $f_{\phi_j}(x)$

Plan for Today

Recap

- Meta-learning problem & black-box meta-learning

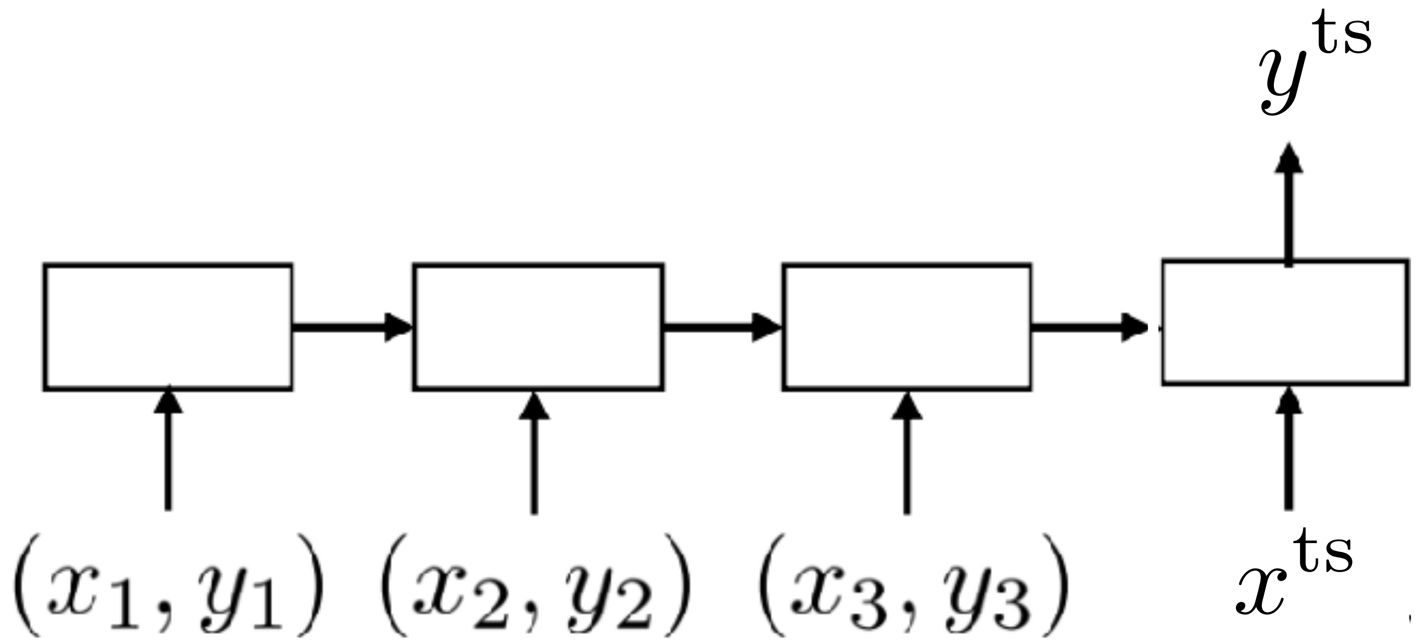
Optimization Meta-Learning } Part of Homework 2!

- Overall approach
- **Compare: optimization-based vs. black-box**
- Challenges & solutions
- Case study of land cover classification (time-permitting)

Optimization vs. Black-Box Adaptation

Black-box adaptation

general form: $y^{\text{ts}} = f_{\text{black-box}}(\mathcal{D}_i^{\text{tr}}, x^{\text{ts}})$



Model-agnostic meta-learning

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

where $\phi_i = \theta - \alpha \nabla_\theta \mathcal{L}(\theta, \mathcal{D}_i^{\text{tr}})$

MAML can be viewed as **computation graph**, with embedded gradient operator

Note: Can mix & match components of computation graph

Learn initialization but replace gradient update with learned network

where $\phi_i = \theta - \alpha$ ~~$\nabla_\theta \mathcal{L}(\theta, \mathcal{D}_i^{\text{tr}})$~~

$f(\theta, \mathcal{D}_i^{\text{tr}}, \nabla_\theta \mathcal{L})$

Ravi & Larochelle ICLR '17
(actually precedes MAML)

This **computation graph view** of meta-learning will come back again!

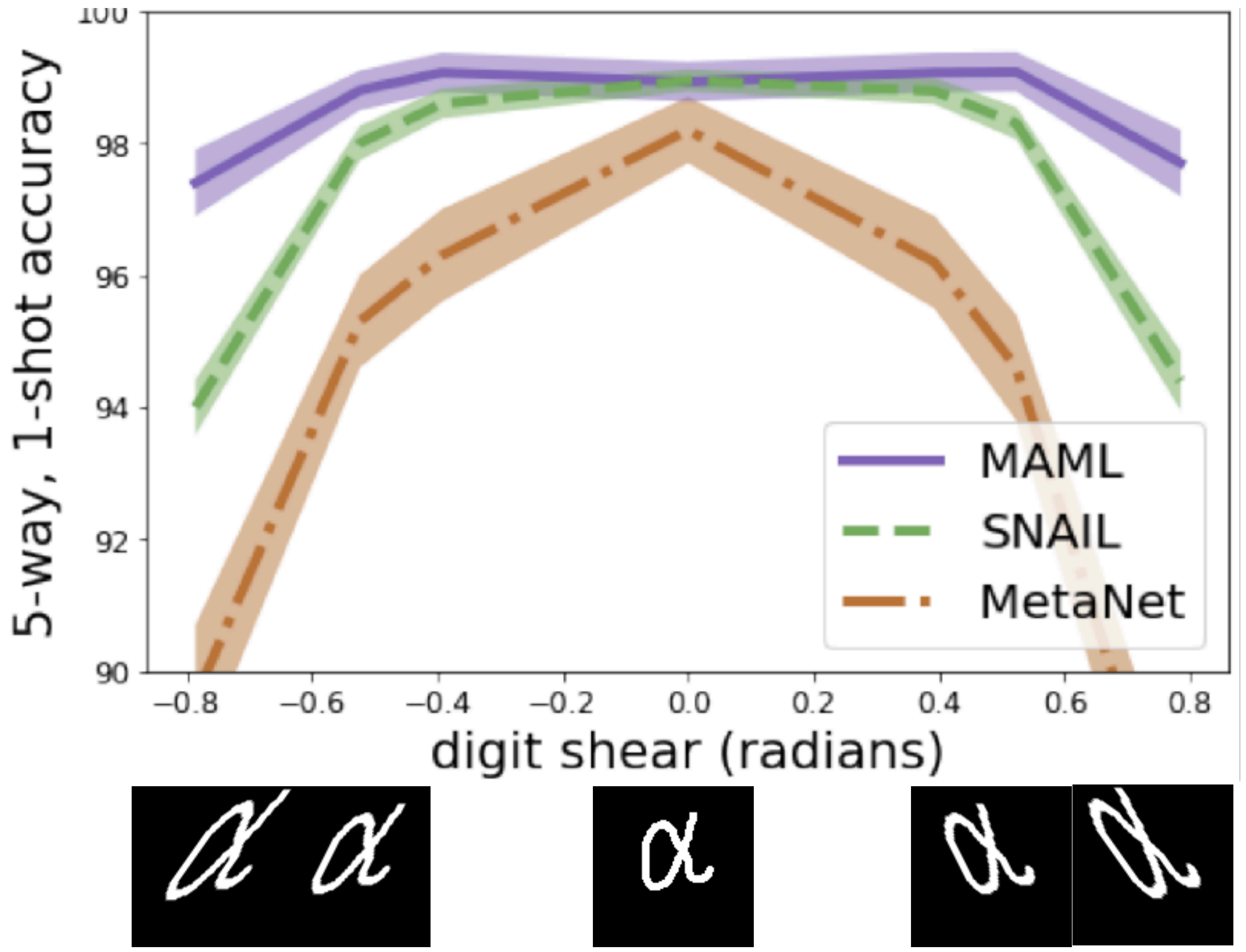
Optimization vs. Black-Box Adaptation

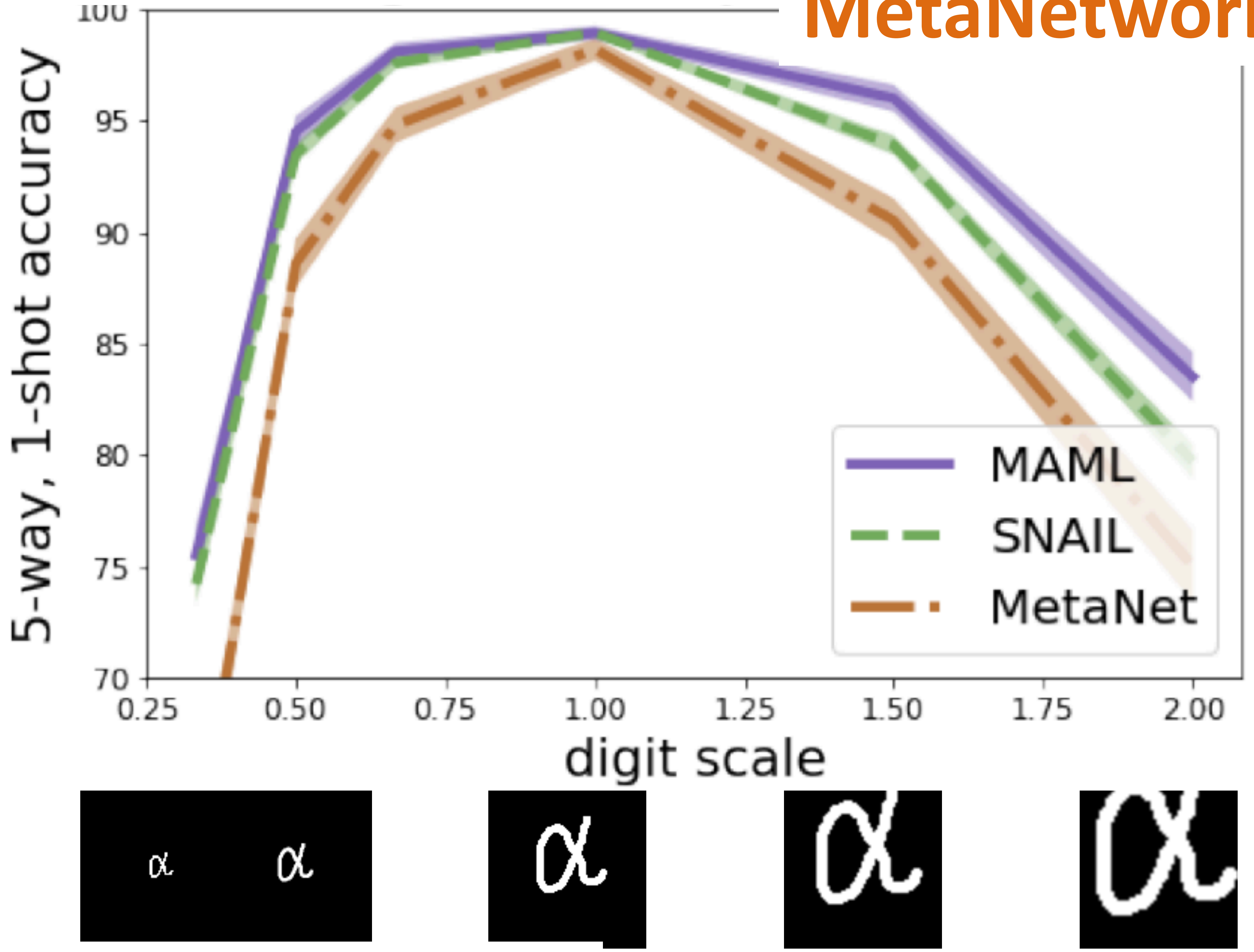
How well can learning procedures generalize to similar, but extrapolated tasks?

Omniglot image classification

MAML SNAIL, MetaNetworks

performance





Does this structure come at a cost?

Finn & Levine ICLR ’18

Black-box adaptation

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

Optimization-based (MAML)

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

Does this structure come at a cost?

For a sufficiently deep network,

MAML function can approximate any function of $\mathcal{D}_i^{\text{tr}}, x^{\text{ts}}$

Finn & Levine, ICLR 2018

Assumptions:

- nonzero α
- loss function gradient does not lose information about the label
- datapoints in $\mathcal{D}_i^{\text{tr}}$ are unique

Why is this interesting?

MAML has benefit of inductive bias without losing expressive power.

Plan for Today

Recap

- Meta-learning problem & black-box meta-learning

Optimization Meta-Learning } Part of Homework 2!

- Overall approach
- Compare: optimization-based vs. black-box
- **Challenges & solutions**
- Case study of land cover classification (time-permitting)

Optimization-Based Adaptation

Challenges. Bi-level optimization can exhibit instabilities.

Idea: Automatically learn inner vector learning rate, tune outer learning rate
(Li et al. Meta-SGD, Behl et al. AlphaMAML)

Idea: Optimize only a subset of the parameters in the inner loop
(Zhou et al. DEML, Zintgraf et al. CAVIA)

Idea: Decouple inner learning rate, BN statistics per-step (Antoniou et al. MAML++)

Idea: Introduce context variables for increased expressive power.
(Finn et al. bias transformation, Zintgraf et al. CAVIA)

Takeaway: a range of simple tricks that can help optimization significantly

Optimization-Based Adaptation

Challenges. Backpropagating through many inner gradient steps is compute- & memory-intensive.

Idea: [Crudely] approximate $\frac{d\phi_i}{d\theta}$ as identity

(Finn et al. first-order MAML '17, Nichol et al. Reptile '18)

Surprisingly works for simple few-shot problems, but (anecdotally) not for more complex meta-learning problems.

Idea: Only optimize the *last layer* of weights.

ridge regression, *logistic regression* (Bertinetto et al. R2-D2 '19)

support vector machine (Lee et al. MetaOptNet '19)

—> leads to a closed form or convex optimization on top of meta-learned features

Idea: Derive meta-gradient using the implicit function theorem

(Rajeswaran, Finn, Kakade, Levine. Implicit MAML '19)

—> compute full meta-gradient ***without differentiating through optimization path***

Optimization-Based Adaptation

Key idea: Acquire ϕ_i through optimization.

Takeaways: Construct *bi-level optimization* problem.

+ positive inductive bias at the start of meta-learning
+ tends to extrapolate better via structure of optimization
+ maximally expressive with sufficiently deep network
+ model-agnostic (easy to combine with your favorite architecture)

- typically requires second-order optimization
- usually compute and/or memory intensive

-> Can be prohibitively expensive for large models

Plan for Today

Recap

- Meta-learning problem & black-box meta-learning

Optimization Meta-Learning } Part of Homework 2!

- Overall approach
- Compare: optimization-based vs. black-box
- Challenges & solutions
- **Case study of land cover classification** (time-permitting)

Case Study

Meta-Learning for Few-Shot Land Cover Classification

Marc Rußwurm[1,*,†], Sherrie Wang[2,3,*], Marco Körner[1], and David Lobell[2]

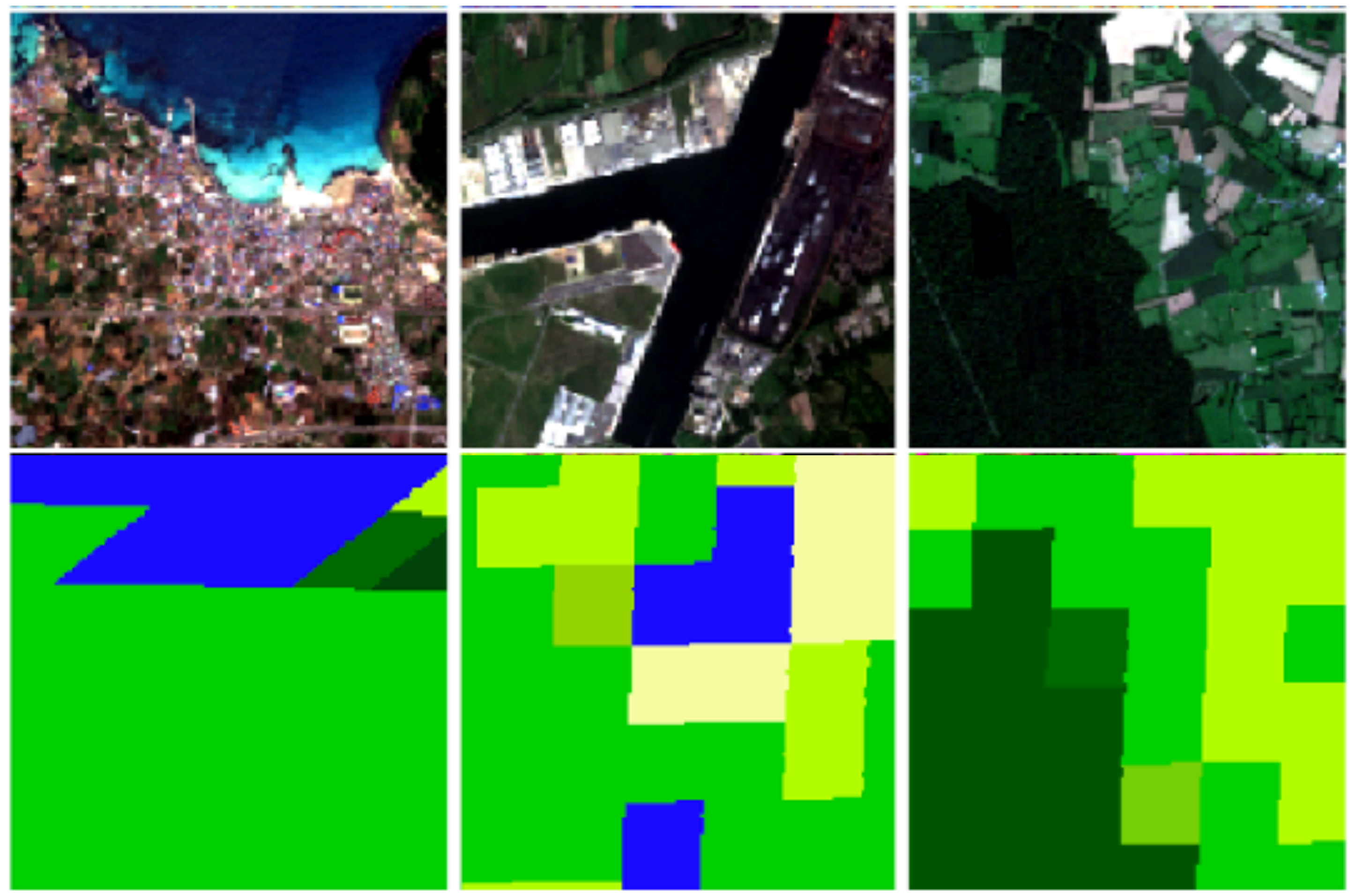
[1]Technical University of Munich, Chair of Remote Sensing Technology
[2]Stanford University, Center on Food Security and the Environment
[3]Stanford University, Institute for Computational and Mathematical Engineering

CVPR 2020 EarthVision Workshop

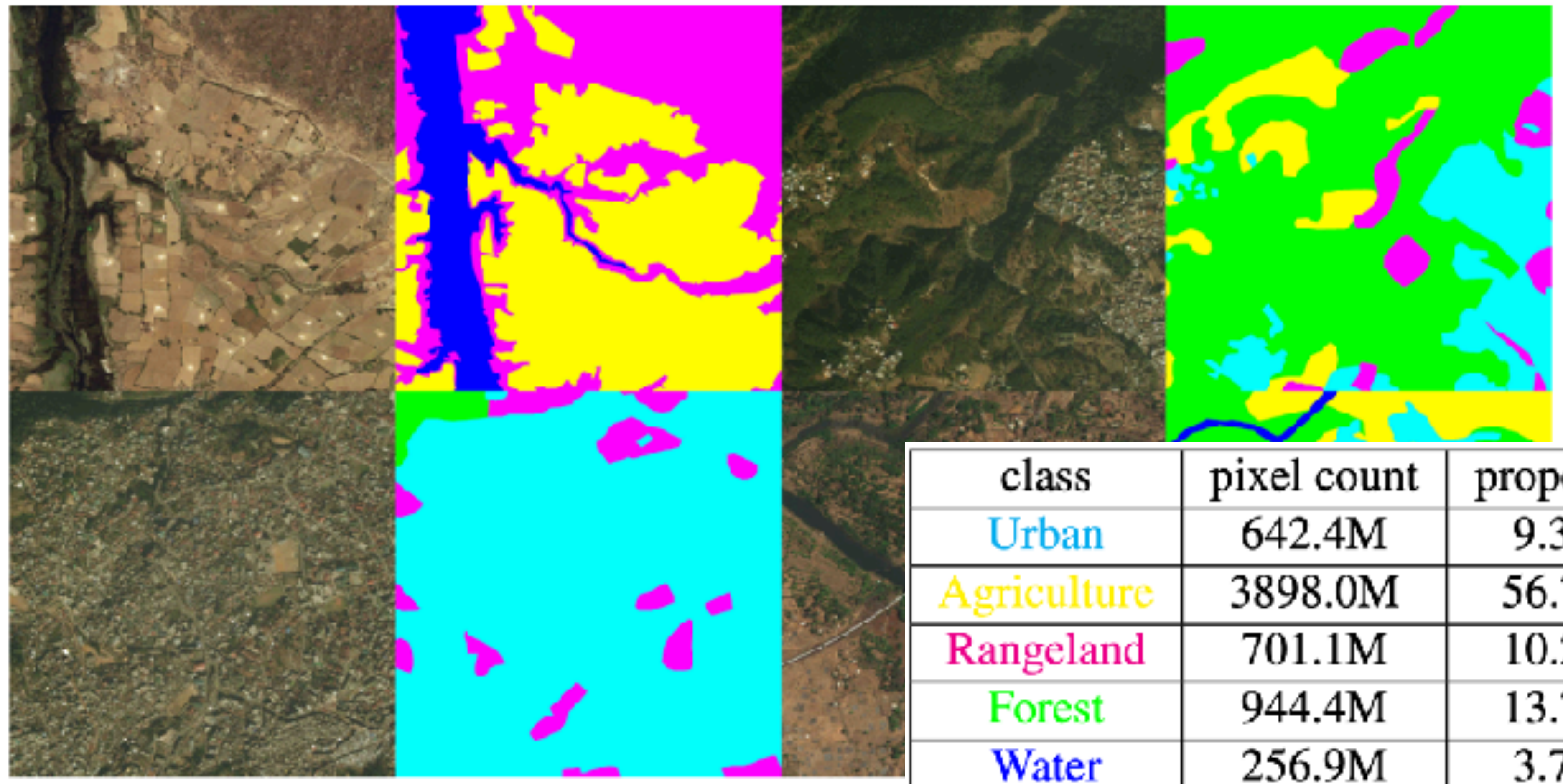
Link: https://arxiv.org/abs/2004.13390

Problem: Map land covering from satellite images

SEN12MS dataset
(Schmitt et al. 2019)



DeepGlobe dataset
(Demir et al. 2018)

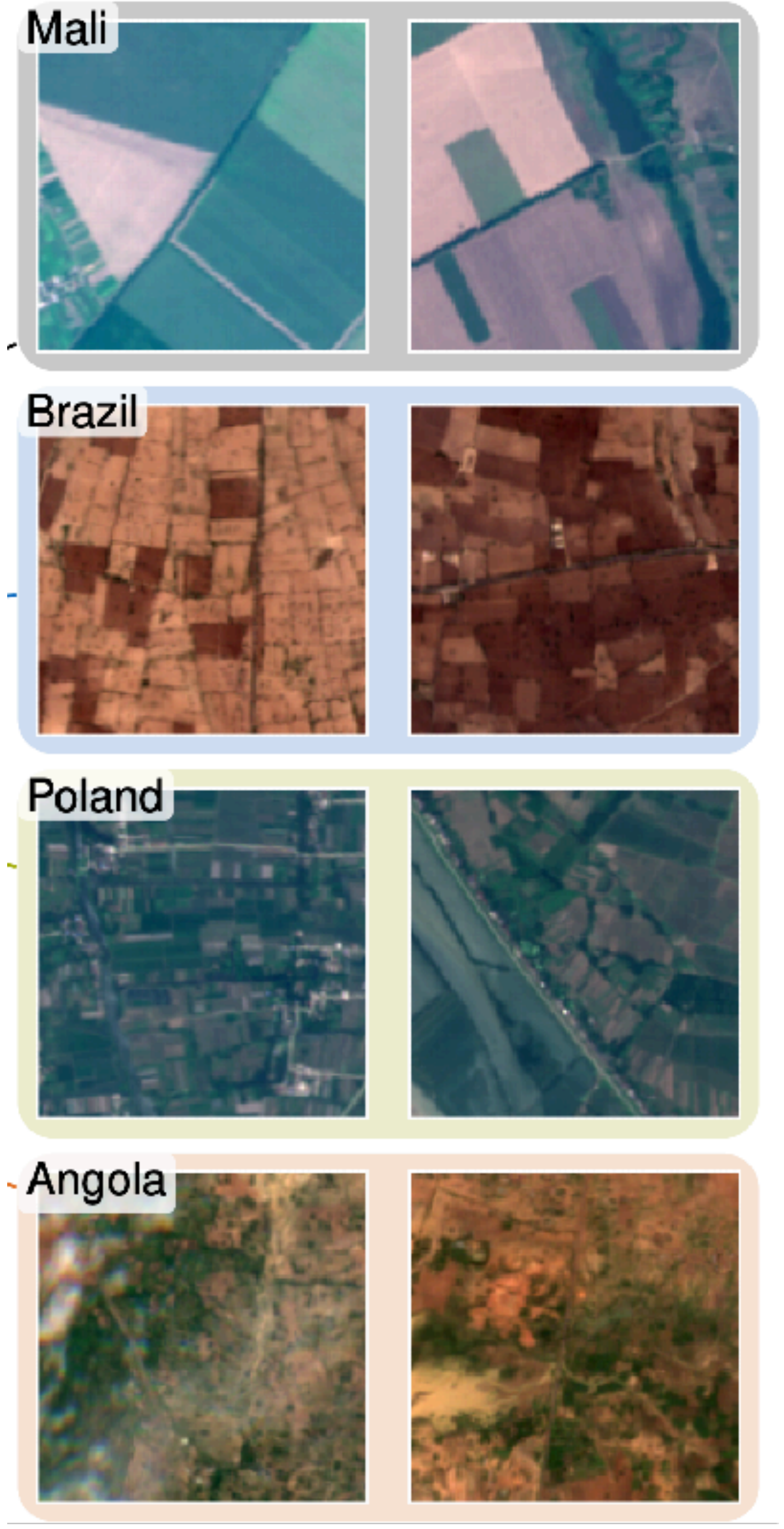


class	pixel count	proportion
Urban	642.4M	9.35%
Agriculture	3898.0M	56.76%
Rangeland	701.1M	10.21%
Forest	944.4M	13.75%
Water	256.9M	3.74%
Barren	421.8M	6.14%
Unknown	3.0M	0.04%

Applications in global urban planning, climate change research

Challenges:
Labeling data is expensive.
Different regions look different & have different land use proportions

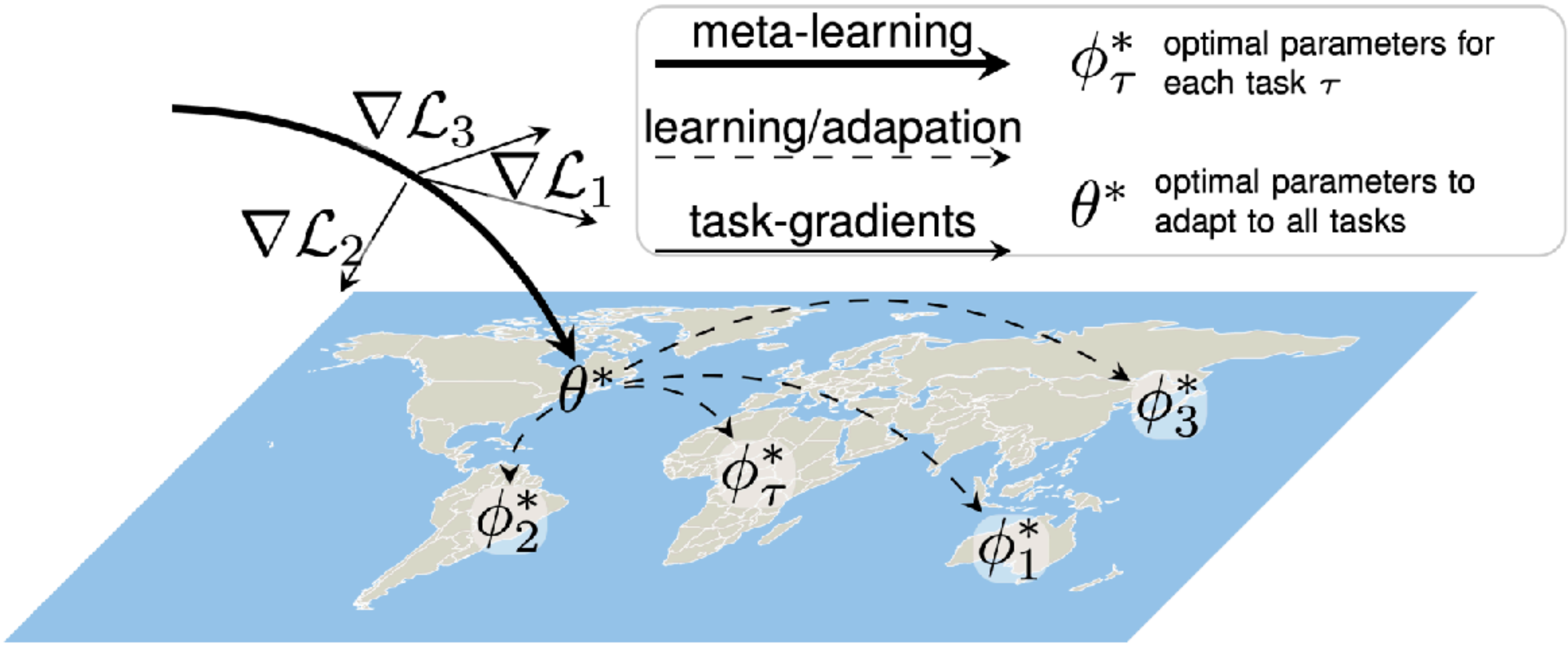
Framing land cover mapping as a meta-learning problem



Croplands from four countries.

Different tasks: different regions of the world

Goal: Segment/classify images from a new region with a small amount of data

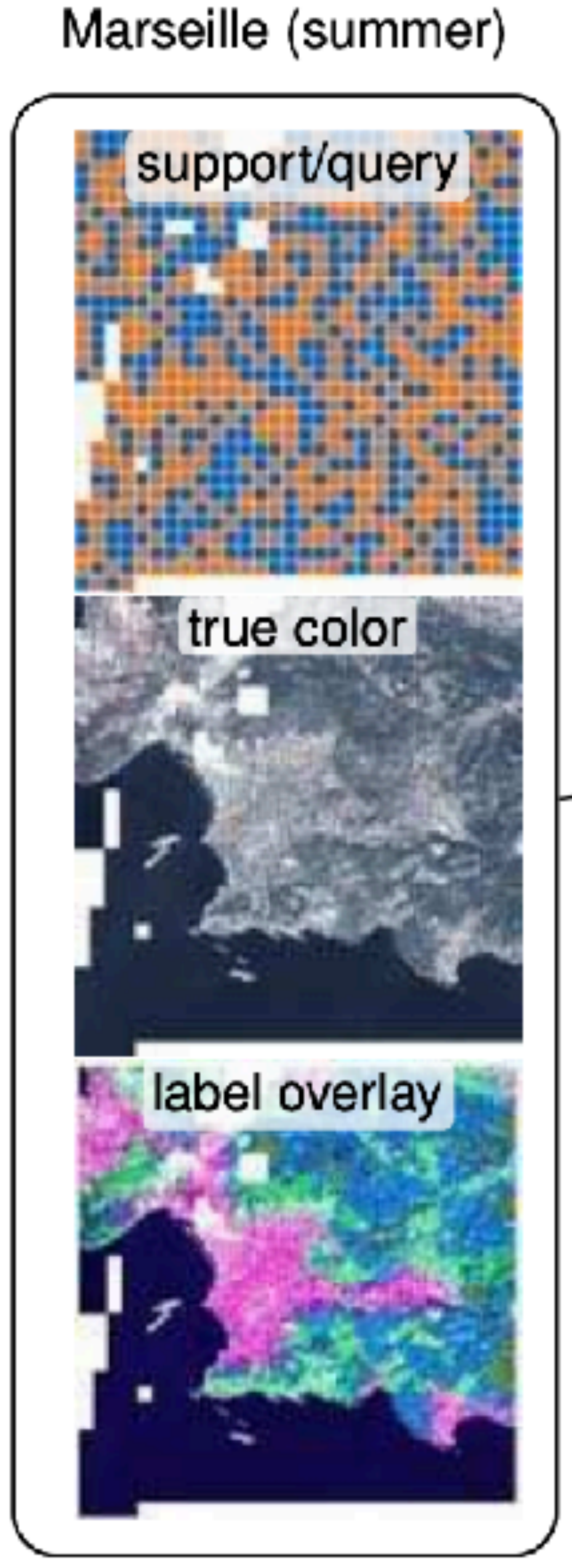


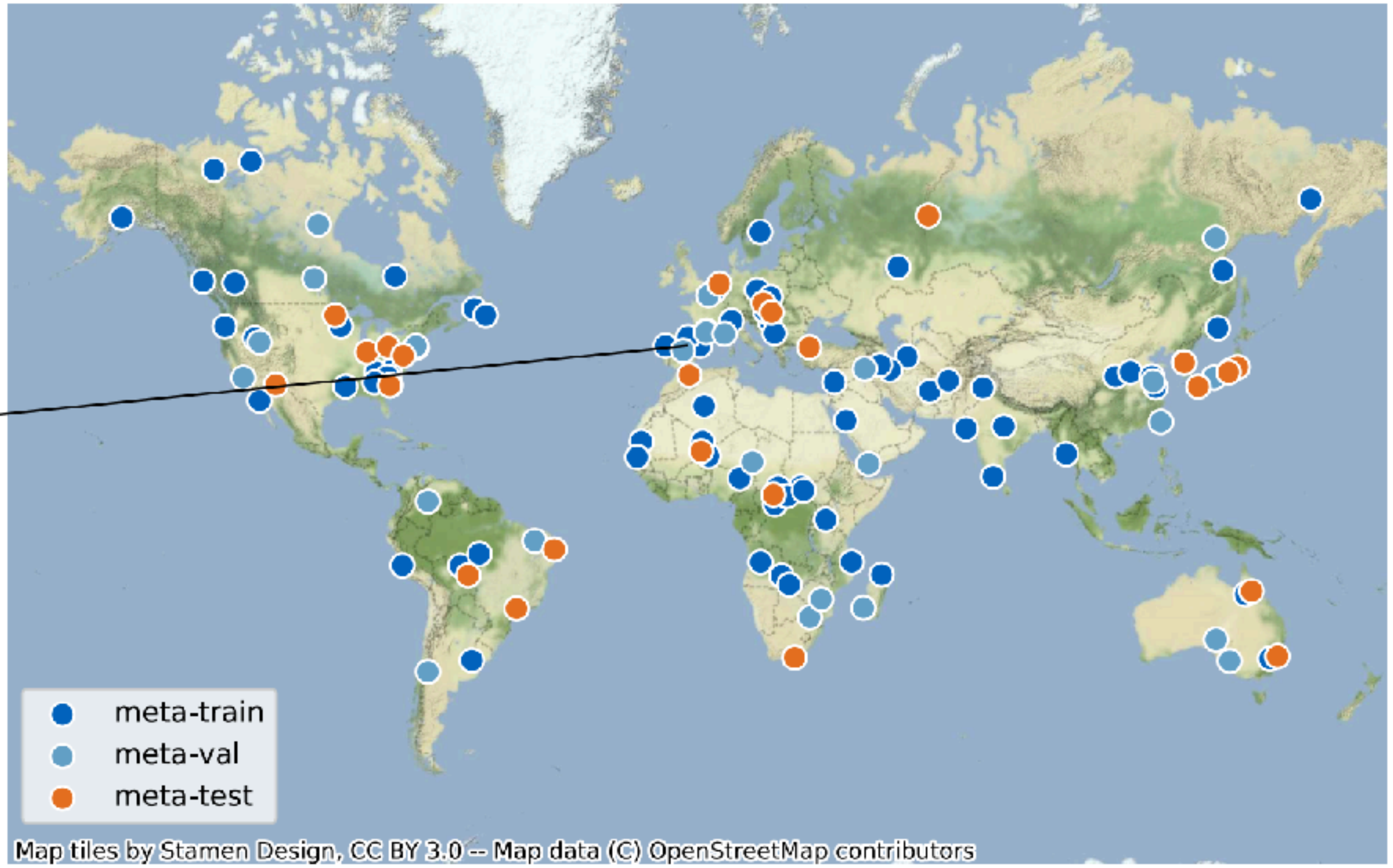
Framing land cover mapping as a meta-learning problem

Goal: Segment/classify images from a new region with a small amount of data

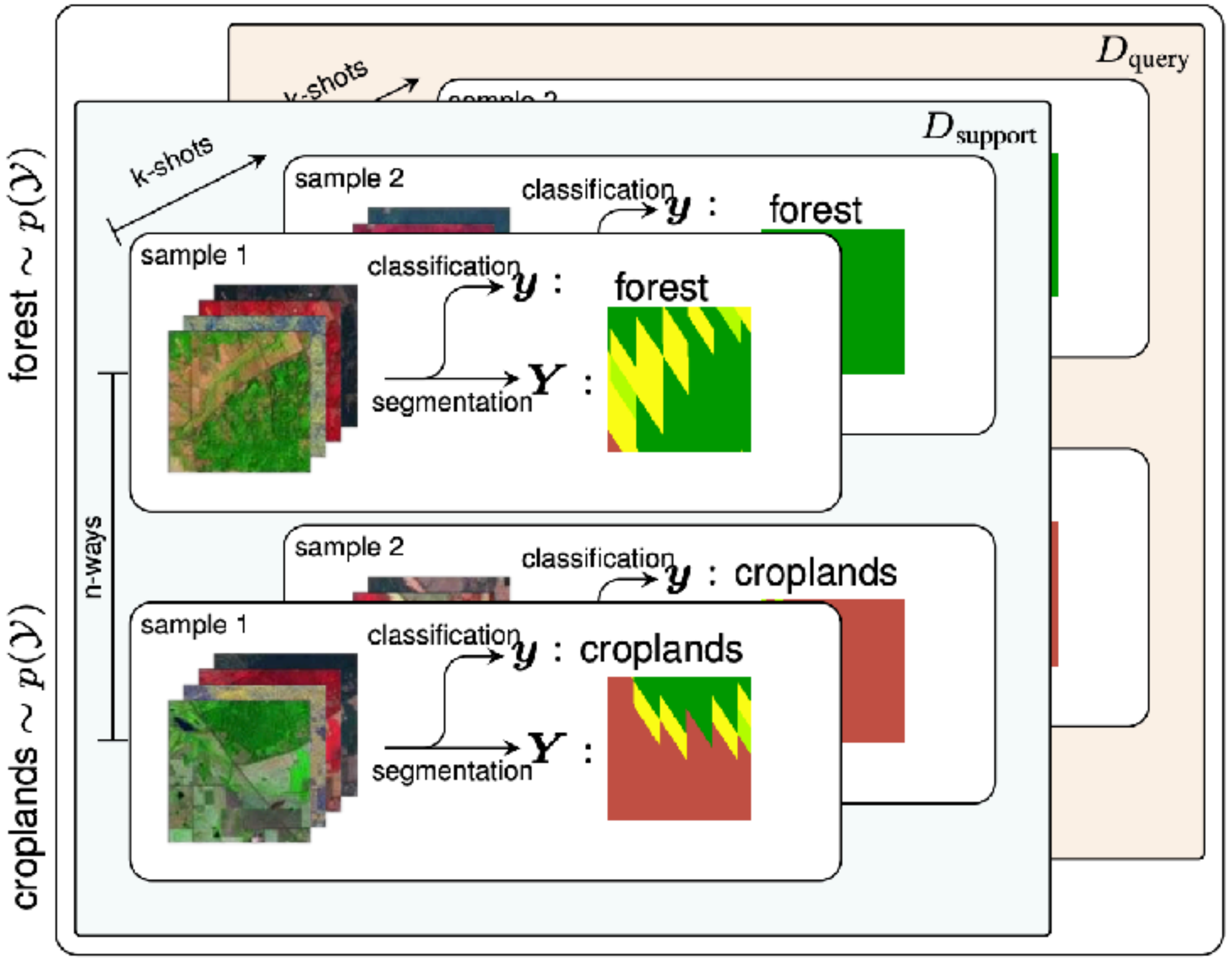
SEN12MS dataset (Schmitt et al. 2019)

Geographic meta-data provided





Example 2-way 2-shot classification task

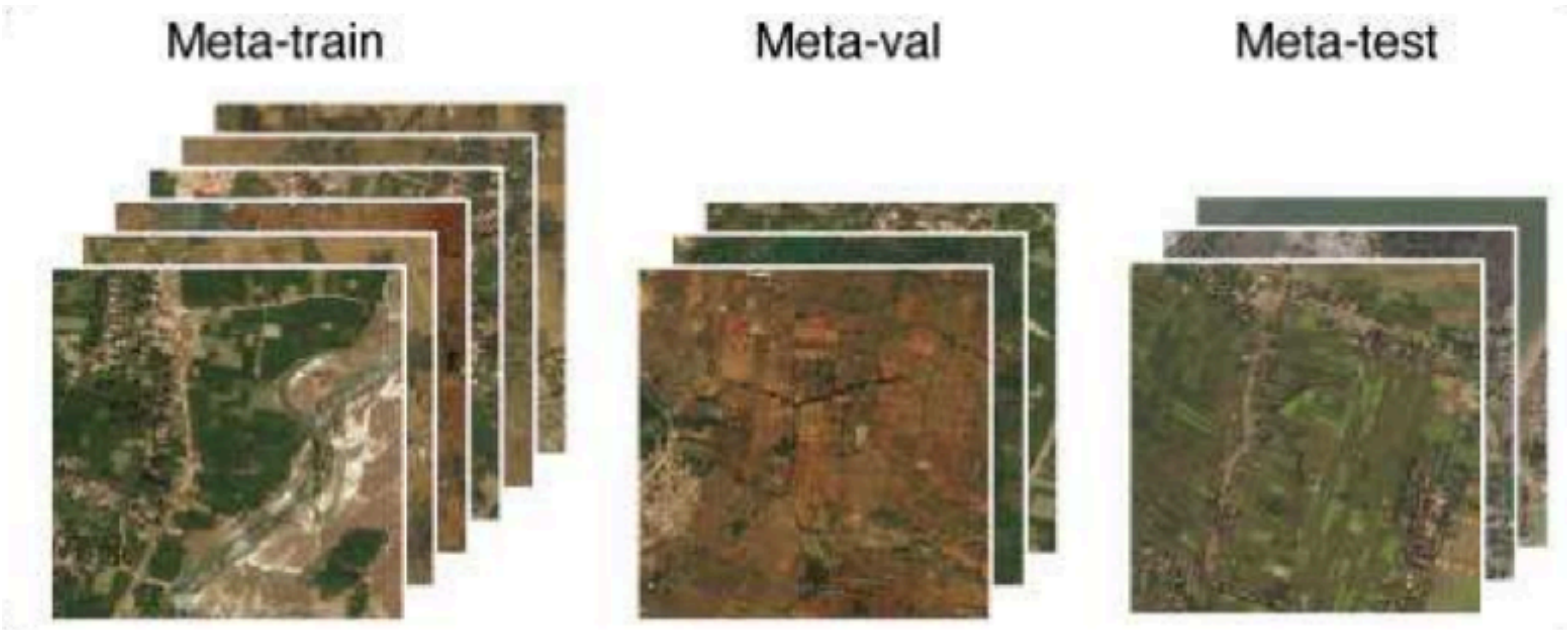


Framing land cover mapping as a meta-learning problem

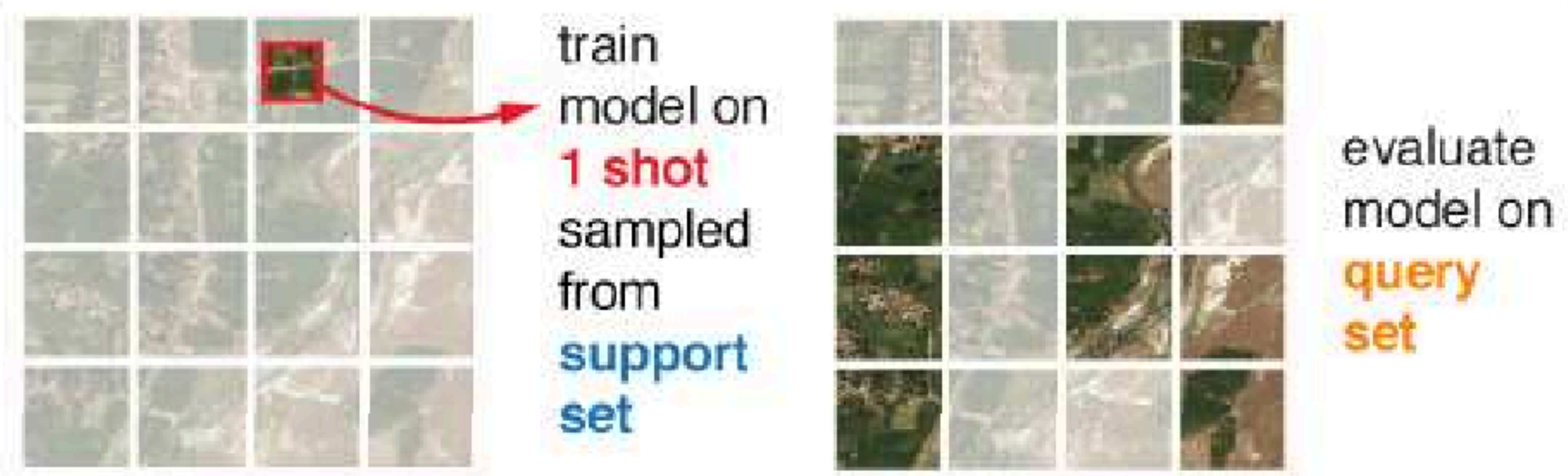
Goal: Segment/classify images from a new region with a small amount of data

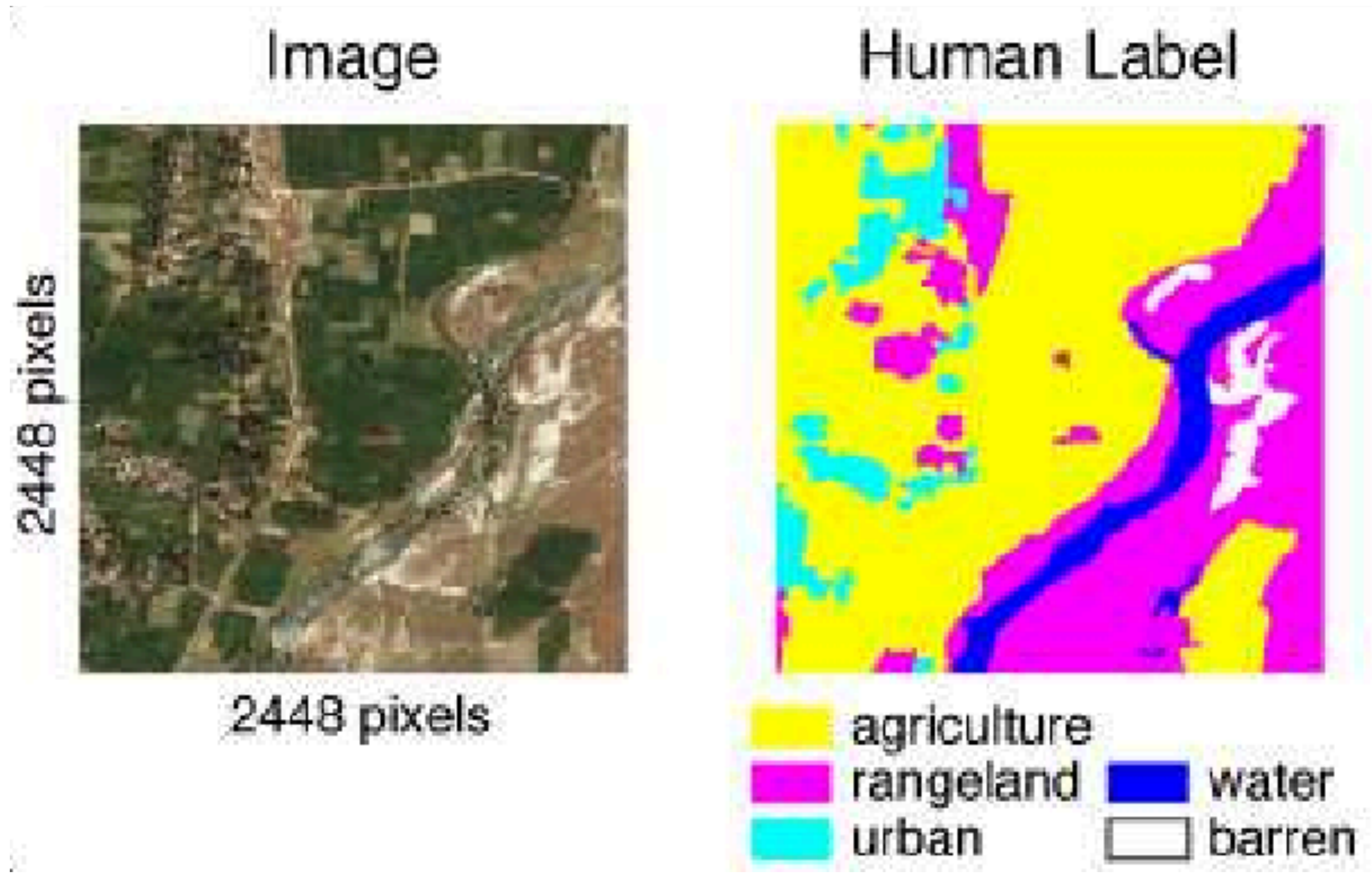
DeepGlobe dataset (Demir et al. 2018)

No geographic metadata, used clustering to guess region



Example 1-shot learning segmentation task.





Evaluation

Meta-training data: $\{\mathscr{D}_1, \ldots, \mathscr{D}_T\}$

Meta-test time: small amount of data from new region: $\mathscr{D}_j^{\mathrm{tr}}$
(meta-test training set / meta-test support set)

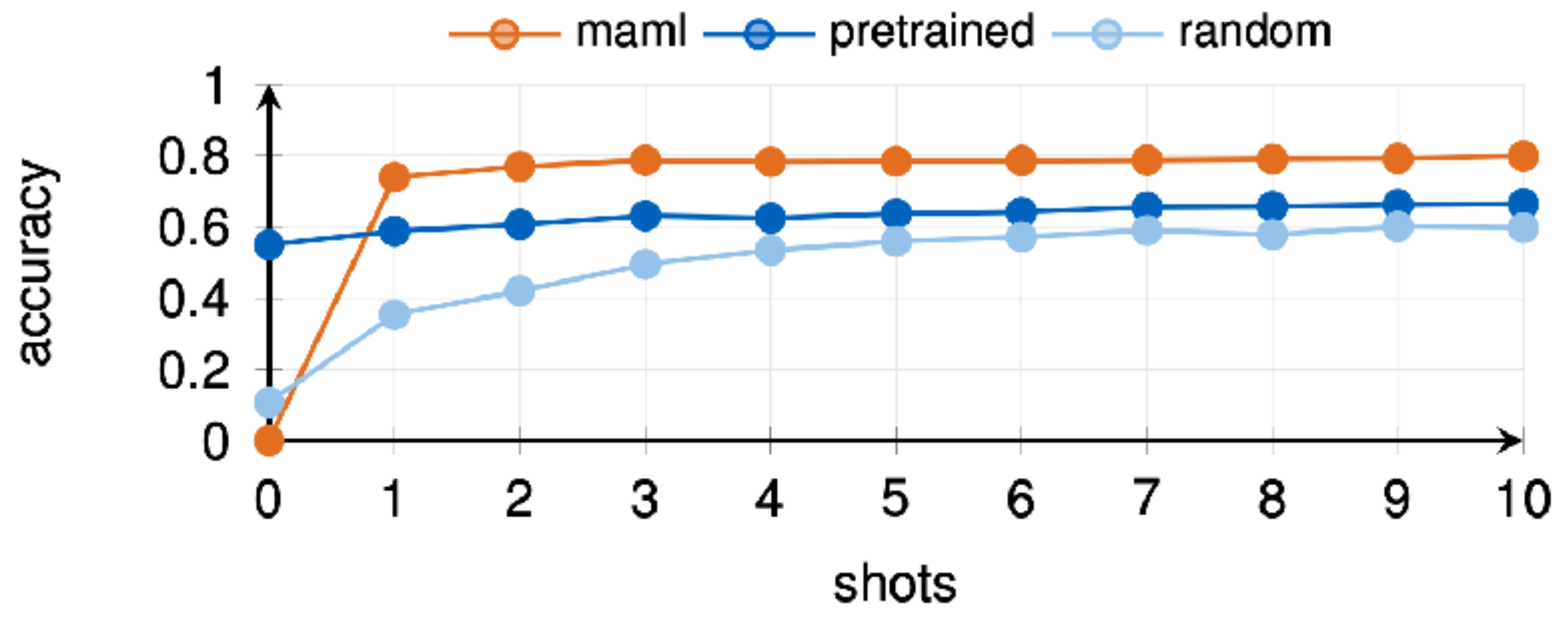
Compare:

Random init: Train from scratch on $\mathscr{D}_j^{\mathrm{tr}}$

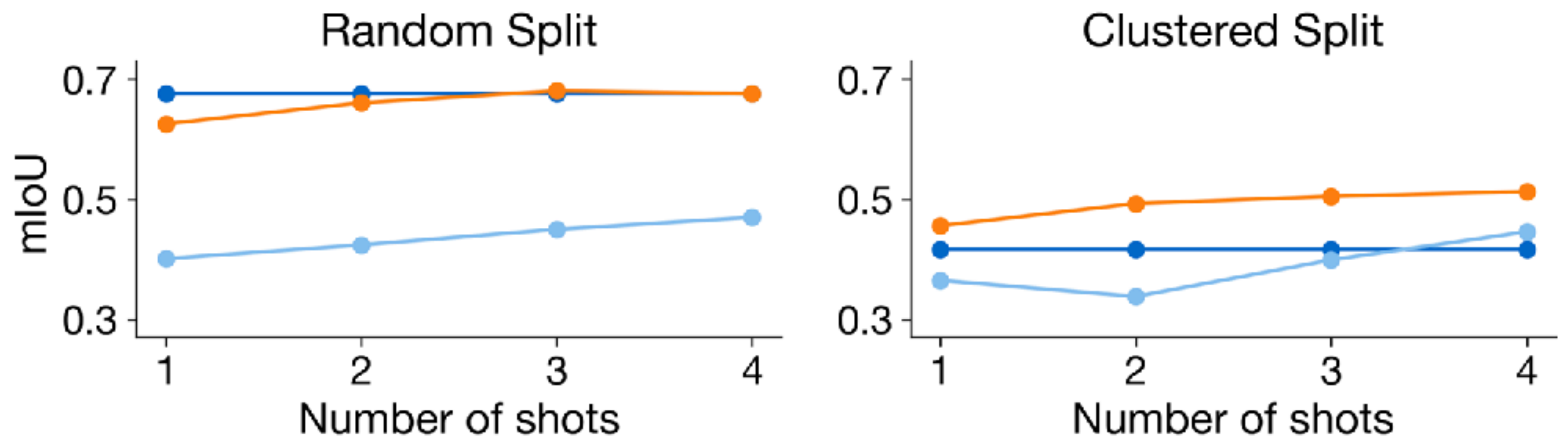
Pre-train on meta-training data $\mathscr{D}_1 \cup \ldots \cup \mathscr{D}_T$, **fine-tune** on $\mathscr{D}_j^{\mathrm{tr}}$

MAML on meta-training data $\{\mathscr{D}_1, \ldots, \mathscr{D}_T\}$, adapt with $\mathscr{D}_j^{\mathrm{tr}}$

SEN12MS dataset



DeepGlobe dataset



More visualizations and analysis in the paper!

Plan for Today

Recap

- Meta-learning problem & black-box meta-learning

Optimization Meta-Learning } **Part of Homework 2!**

- Overall approach
- Compare: optimization-based vs. black-box
- Challenges & solutions
- Case study of land cover classification (time-permitting)

Goals for by the end of lecture:

- Basics of optimization-based meta-learning techniques (& how to implement)
- Trade-offs between black-box and optimization-based meta-learning

Roadmap for upcoming lectures

Monday: Non-parametric few-shot learners, comparison of approaches

Weds & Next Monday: Unsupervised pre-training for few-shot learning

Following lectures: Advanced meta-learning topics (e.g. memorization, large-scale meta-optimization)

Course Reminders

Project group form due **tonight**.

(for assigning project mentors)

Homework 1 due **Monday**

Homework 2 out **today**

Tutorial session **tomorrow 4:30-5:20 pm** on MAML.