# More about Meta Learning

Hung-yi Lee

# Prerequisite

https://youtu.be/xoastiYx9JU

https://youtu.be/Q68Eh-wm1Ts

# Outline

- Meta Learning vs. Self-supervised Learning
- Meta Learning vs. Domain Generalization
- Meta Learning vs. Knowledge Distillation
- Meta Learning vs. Life-long Learning

# Meta Learning vs. Self-supervised Learning

# Meta Learning
# vs. Self-supervised Learning

Self-supervised Learning (BERT and pals)

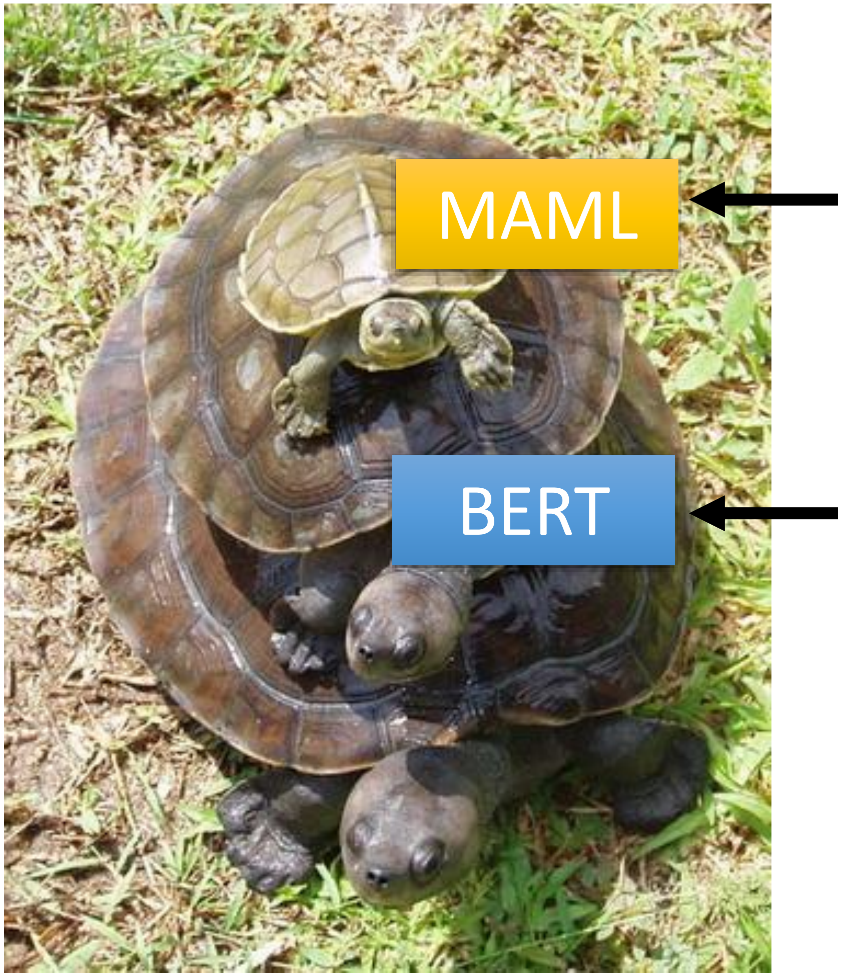
Learn to Init (MAML family)

# Meta Learning vs. Self-supervised Learning

- MAML learns the initialization parameter $\phi$ by gradient descent
- What is the initialization parameter $\phi^0$ for $\phi$?

BERT can serve as $\phi^0$

# Meta Learning
# vs. Self-supervised Learning

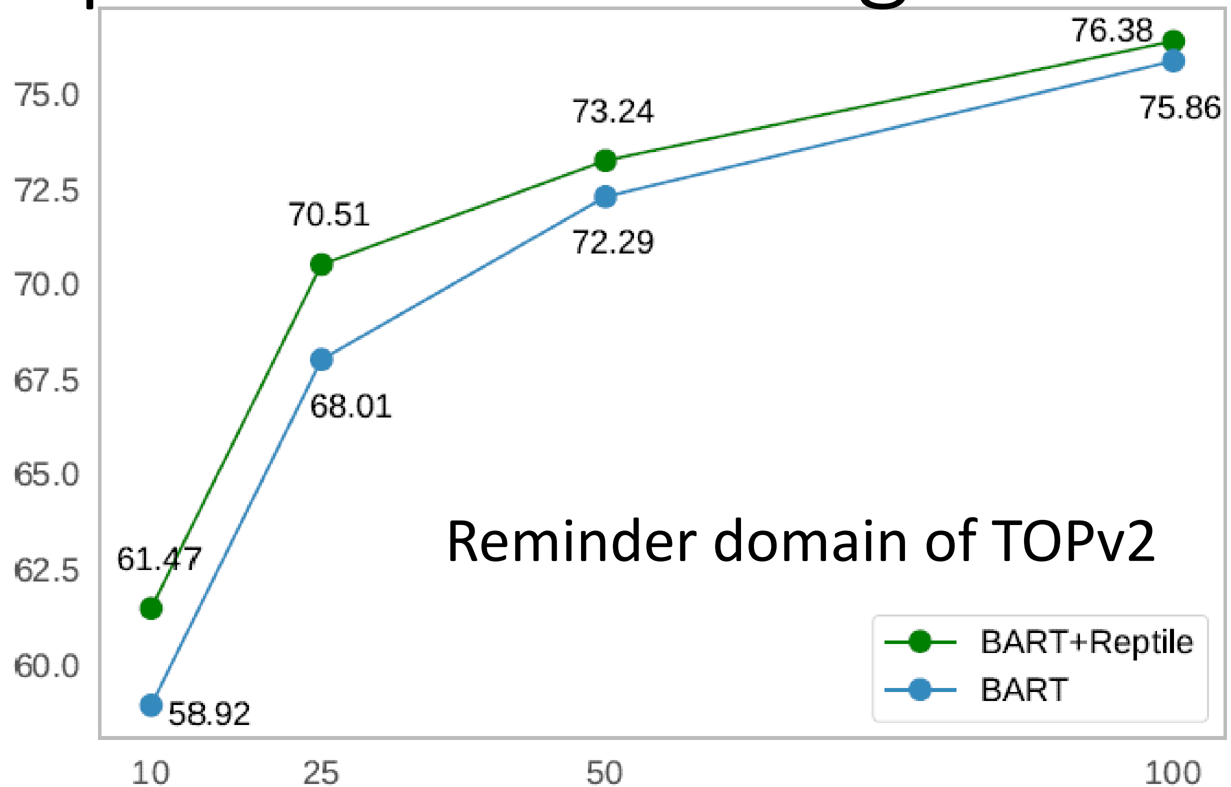



- Learn to achieve good performance on training tasks.
- Leverage training tasks.
- The self-supervised objectives are different from downstream tasks.

There is a "*learning gap*".

- Utilize a large amount of unlabeled data

# Meta Learning vs. Self-supervised Learning

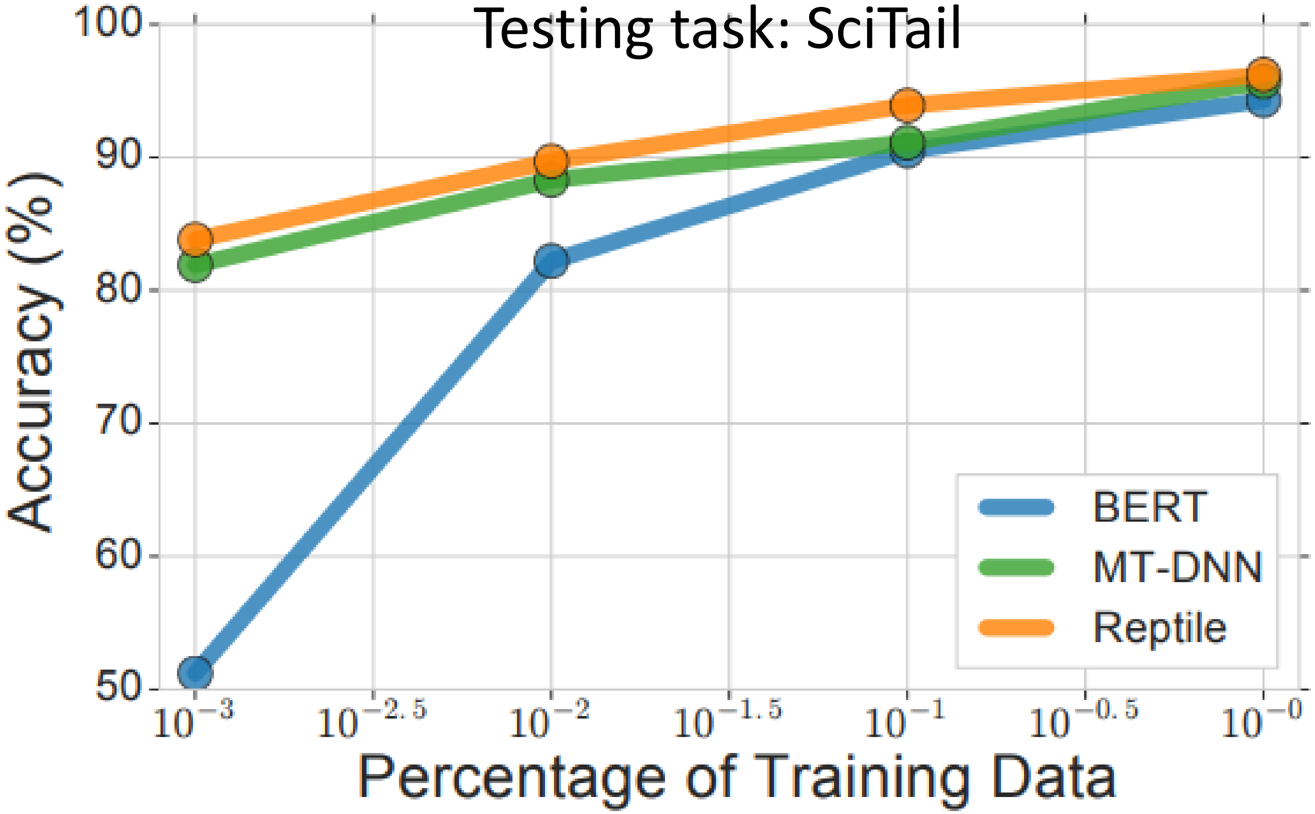



SPIS = samples per intent and slot

Xilun Chen, Asish Ghoshal, Yashar Mehdad, Luke Zettlemoyer, Sonal Gupta, Low-Resource Domain Adaptation for Compositional Task-Oriented Semantic Parsing, EMNLP, 2020

# Meta Learning vs. Self-supervised Learning

Testing task: SciTail




Zi-Yi Dou, Keyi Yu, Antonios Anastasopoulos, Investigating Meta-Learning Algorithms for Low-Resource Natural Language Understanding Tasks, EMNLP 2019

| Work | Method | How to Initailize the Initailization |
|---|---|---|
| (Bansal et al., 2020a) | LEOPARD | BERT |
| (Li et al., 2020a) | MAML | Word Embedding |
| (Park et al., 2021) | MAML | XLM |
| (Gu et al., 2018) | FOMAML | Word Embedding |
| (Langedijk et al., 2021) | FOMAML | mBERT |
| (Chen et al., 2020b) | Reptile | BART |
| (Huang et al., 2020a) | MAML | BERT |
| (Wang et al., 2021b) | Propose a new method based on Reptile | Word Embedding |
| (Dingliwal et al., 2021) | Reptile | RoBERTa |
| (Qian and Yu, 2019) | MAML | Word Embedding |
| (Qian et al., 2021) | MAML | Word Embedding |
| (Madotto et al., 2019) | MAML | Word Embedding |
| (Dai et al., 2020) | MAML | - |
| (Hsu et al., 2020) | FOMAML | Multilingual ASR |
| (Xiao et al., 2021) | MAML/FOMAML/Reptile | - |
| (Winata et al., 2020b) | MAML | Pretrain by Supervised Learning |
| (Klejch et al., 2019) | FOMAML | - |
| (Huang et al., 2021) | MAML/FOMAML | - |
| (Indurthi et al., 2020) | FOMAML | - |
| (Winata et al., 2020a) | FOMAML | - |
| (Wu et al., 2021b) | MAML | Pretrain by Multi-task Learning |
| (Ke et al., 2021) | MAML | BERT |
| (Xia et al., 2021) | MetaXL | mBERT/XLM-R |
| (Dou et al., 2019) | MAML/FOMAML/Reptile | BERT |
| (Obamuyide and Vlachos, 2019b) | FOMAML | Word Embedding |
| (Lv et al., 2019) | MAML | - |
| (Holla et al., 2020) | FOMAML/Proto(FO)MAML | Word Embedding/ELMo/BERT |
| (Huang et al., 2020b) | MAML | Word Embedding |
| (Mi et al., 2019) | MAML | - |
| (Wang et al., 2021a) | DG-MAML | BERT |
| (Conklin et al., 2021) | DG-MAML | - |
| (M'hamdi et al., 2021) | MAML | mBERT |
| (Nooralahzadeh et al., 2020) | MAML | BERT/mBERT/XLM-R |
| (Garcia et al., 2021) | MAML | mBERT |
| (van der Heijden et al., 2021) | FOMAML/Reptile/Proto(FO)MAML | XLM-R |
| (Bansal et al., 2020b) | LEOPARD | BERT |
| (Murty et al., 2021) | FOMAML | BERT |
| (Hua et al., 2020) | Reptile | - |
| (Yan et al., 2020) | MAML | BERT/RoBERTa |
| (Wang et al., 2019b) | Reptile | - |
| (Bose et al., 2020) | Meta-Graph | - |

https://arxiv.org/abs/2205.01500

# Meta Learning vs. Knowledge Distillation

# Knowledge Distillation

Learning target

“cat”: 0.8, “dog”: 0.2 …

Cross-entropy minimization

?

Teacher Net (Large)

Student Net (Small)

The teacher is not optimized for teaching.

Source of results: https://arxiv.org/abs/2202.07940

# Knowledge Distillation

*Table 1.* **Knowledge distillation is not compatible with strong augmentation.** (Das et al., 2020; Cui & Yan, 2021)

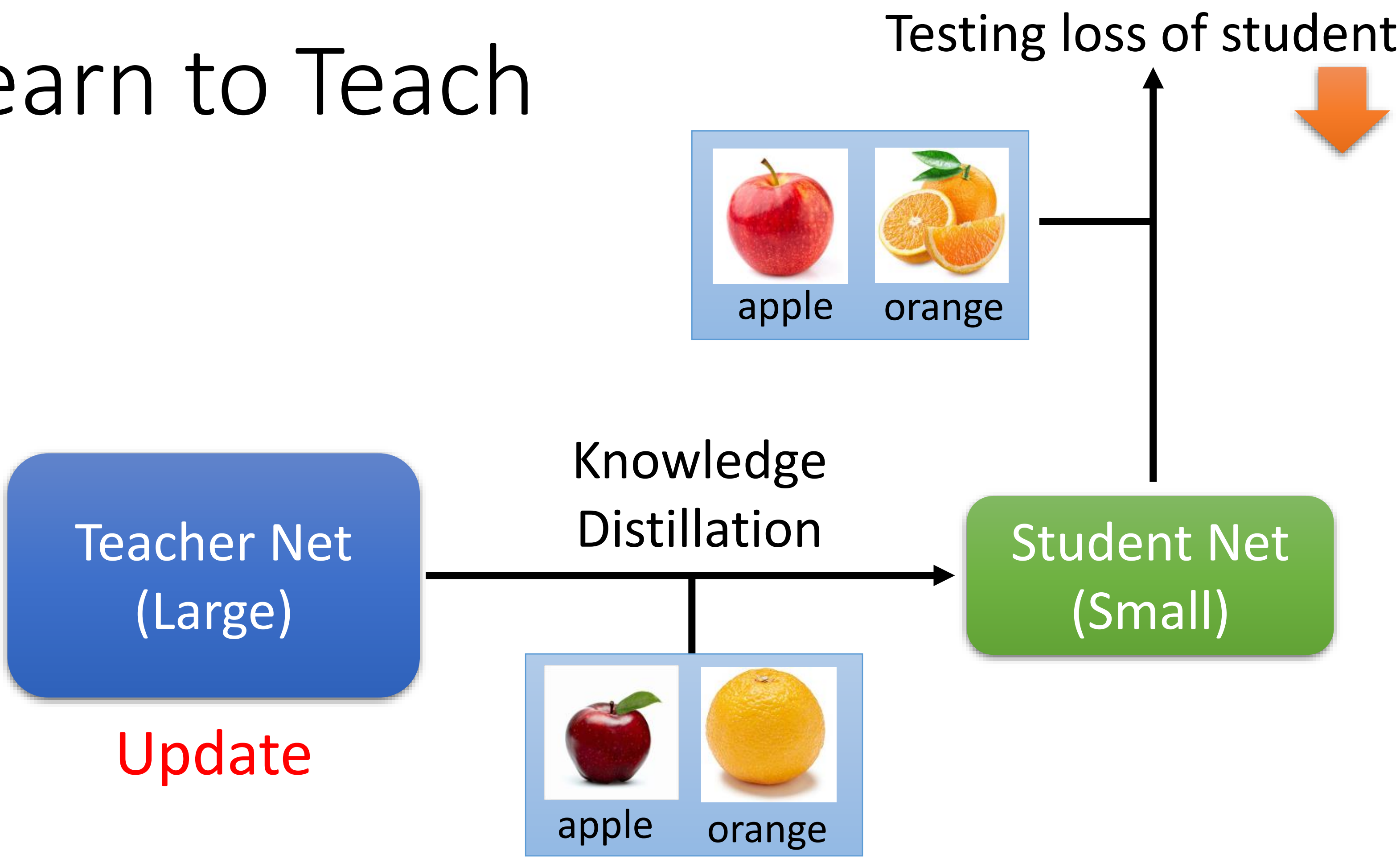
| Augmentation | Teacher acc. | Student acc. |
|---|---|---|
| Normal | 73.3 | **70.4** |
| LabelSmooth (Szegedy et al., 2016) | 73.3 | 68.1 |
| Mixup (Zhang et al., 2017) | 74.3 | 67.9 |
| CutMix (Yun et al., 2019) | **76.0** | 68.8 |

*Table 2.* **Knowledge distillation suffers from teacher-student gap.** (Mirzadeh et al., 2020; Cho & Hariharan, 2019)

| Teacher | Teacher acc. | Student (ResNet20) acc. |
|---|---|---|
| ResNet32 (He et al., 2016) | 71.0 | **70.8** |
| ResNet56 (He et al., 2016) | 73.3 | 70.4 |
| ResNet110 (He et al., 2016) | 75.4 | 70.5 |
| ResNet164 (He et al., 2016) | **75.8** | 69.9 |

Can the teacher-network “learn to teach”?

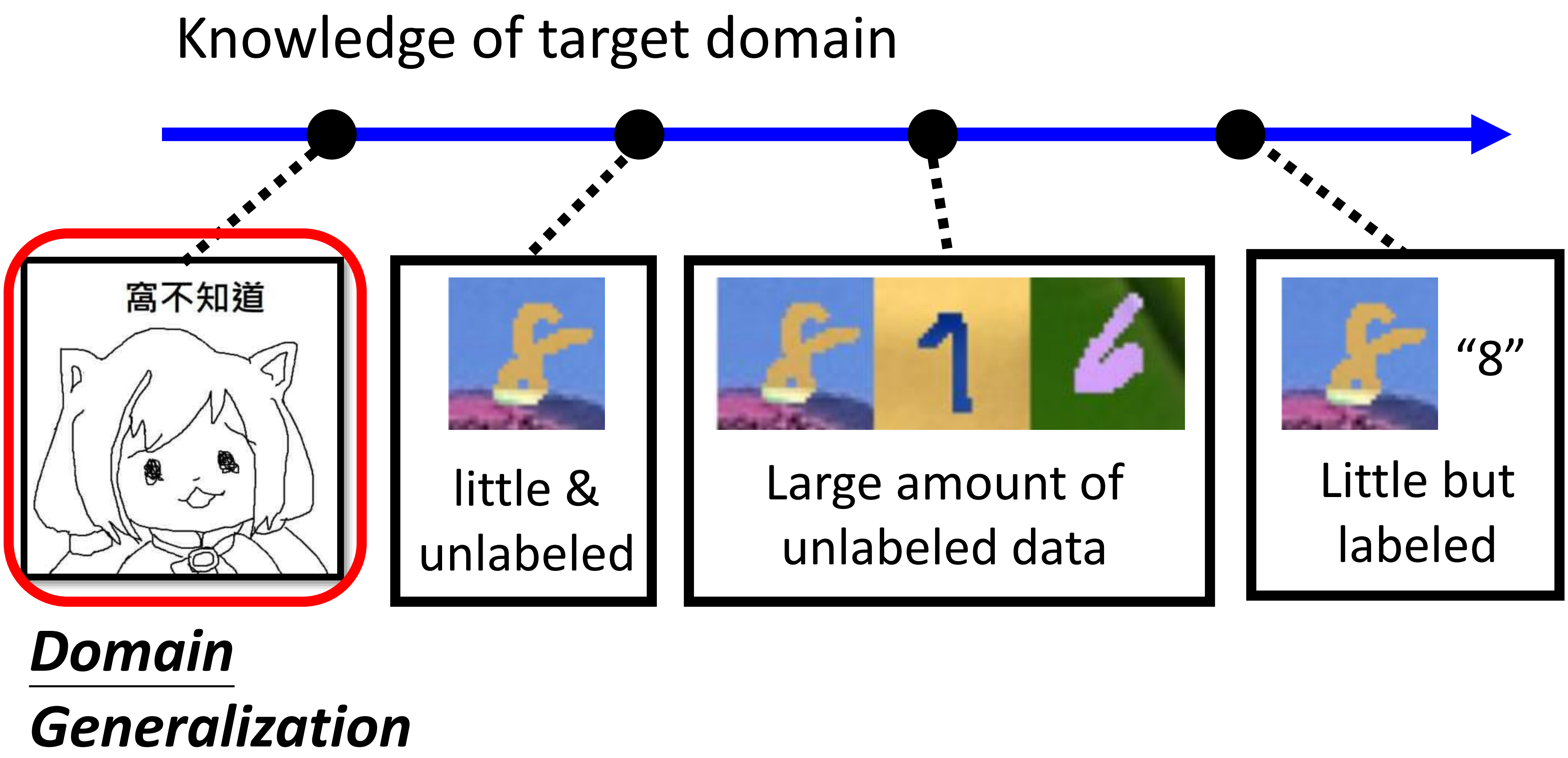
# Learn to Teach



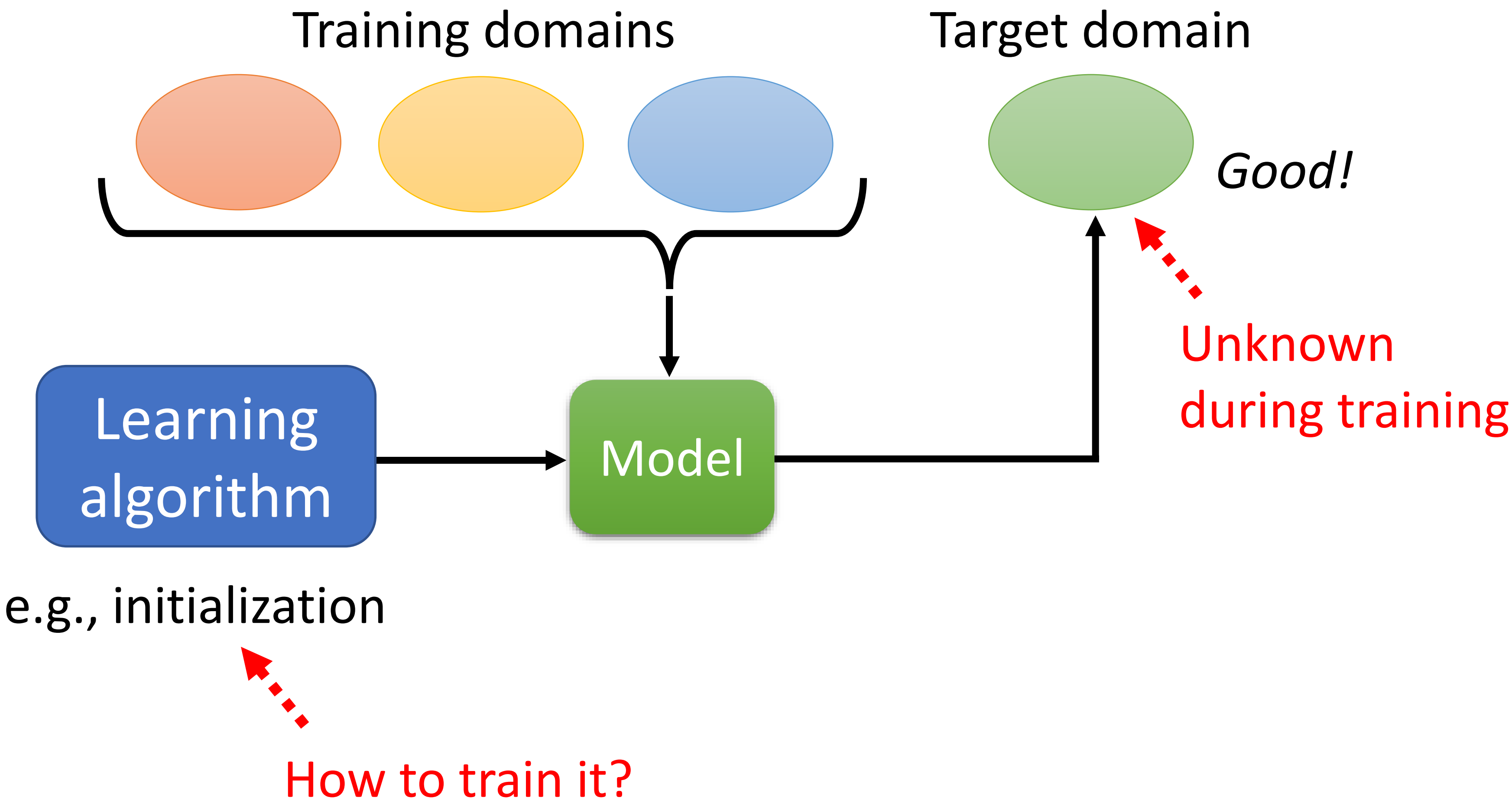

Wangchunshu Zhou, Canwen Xu, Julian McAuley, BERT Learns to Teach: Knowledge Distillation with Meta Learning, ACL, 2022
Jihao Liu, Boxiao Liu, Hongsheng Li, Yu Liu, Meta Knowledge Distillation, arXiv, 2022

# Meta Learning vs. Domain Adaptation

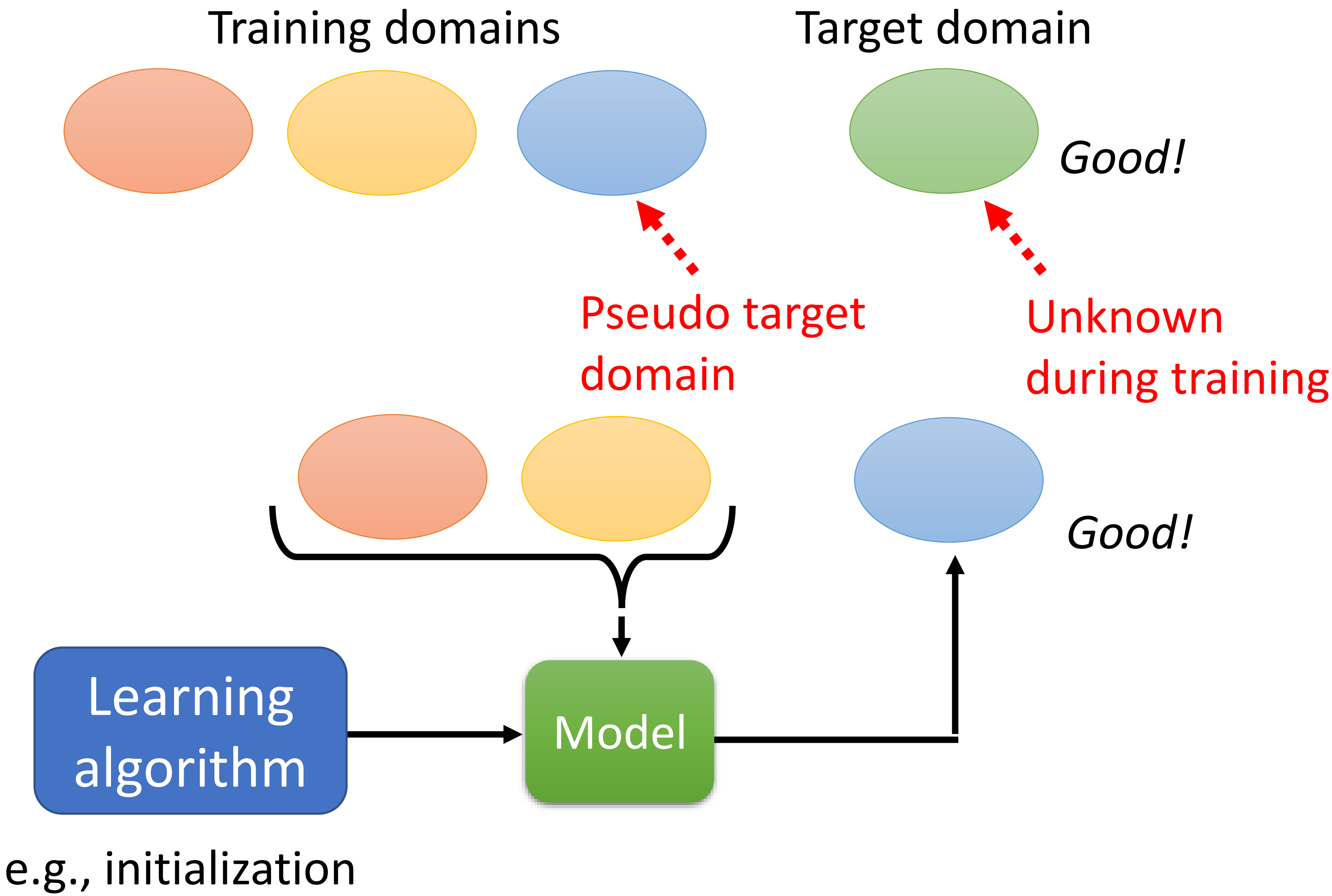
# Domain Adaptation



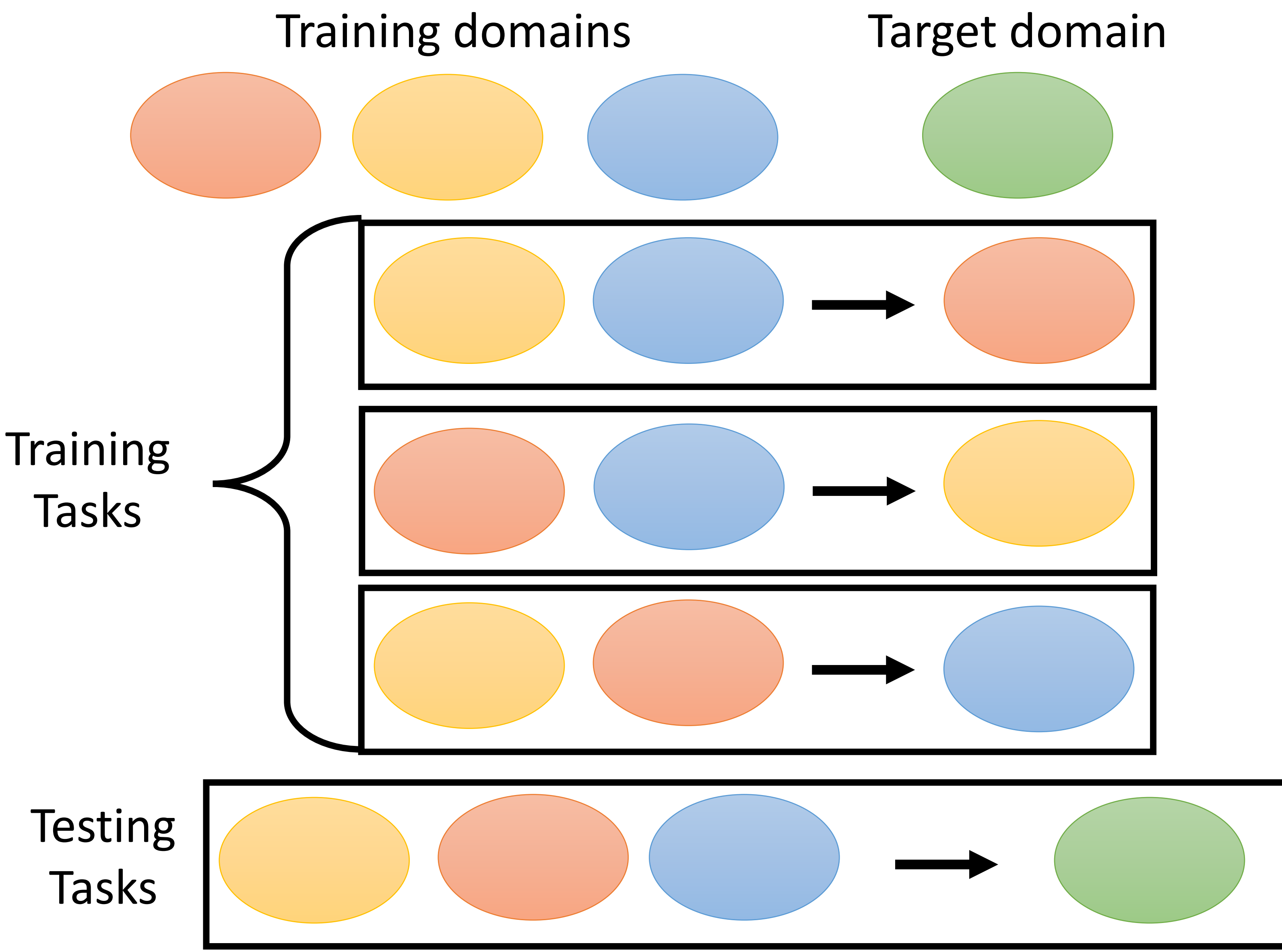
# *Meta Learning for Domain Generalization*



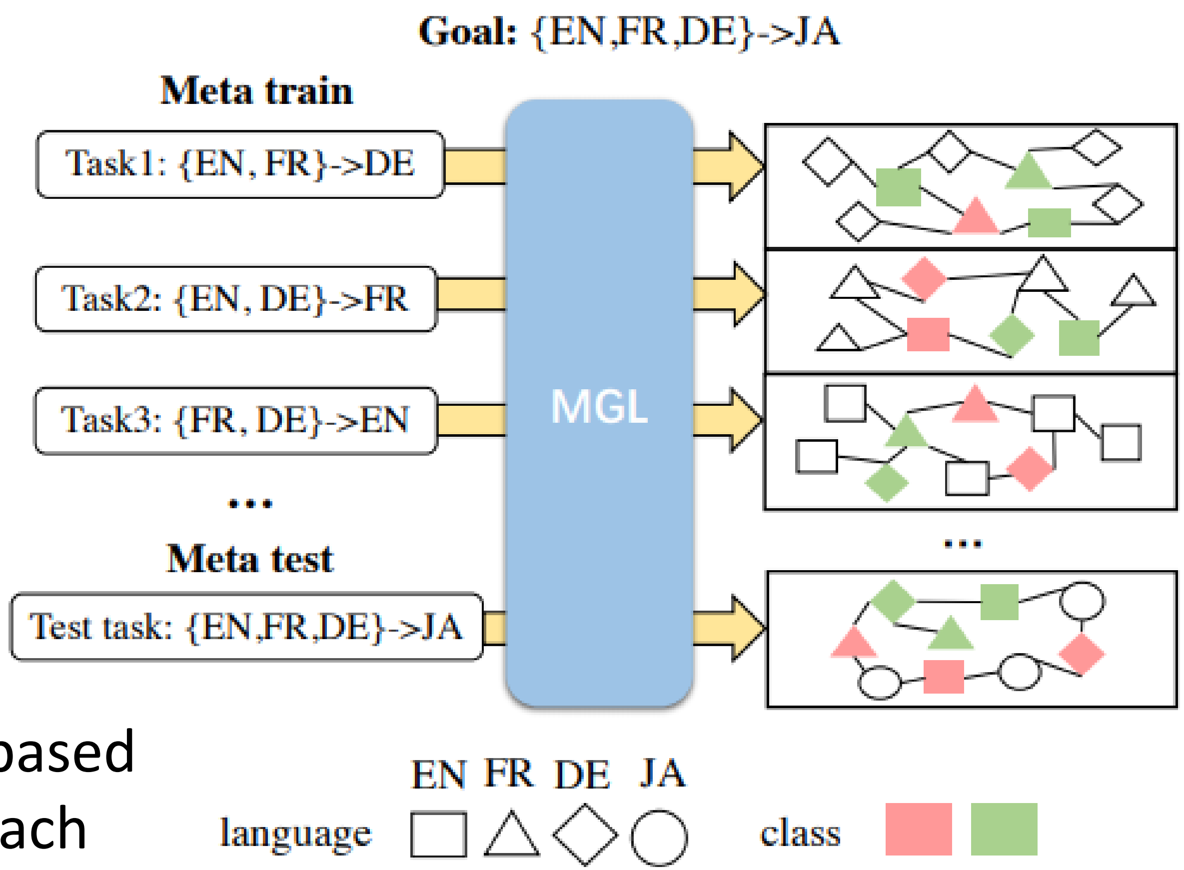
# *Meta Learning for Domain Generalization*



# ***Meta Learning for Domain Generalization***



# Example – Text Classification




Metric-based Approach

Zheng Li, Mukul Kumar, William Headden, Bing Yin, Ying Wei, Yu Zhang, Qiang Yang, Learn to Cross-lingual Transfer with Meta Graph Learning Across Heterogeneous Languages, EMNLP, 2020

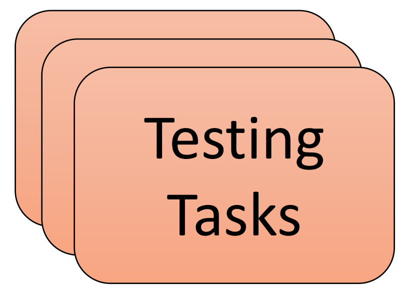
# ***Problem of another level ……***

- The training examples and testing examples may have different distributions.

Training Examples

Testing Examples

- The training tasks and testing tasks can also have different distributions.

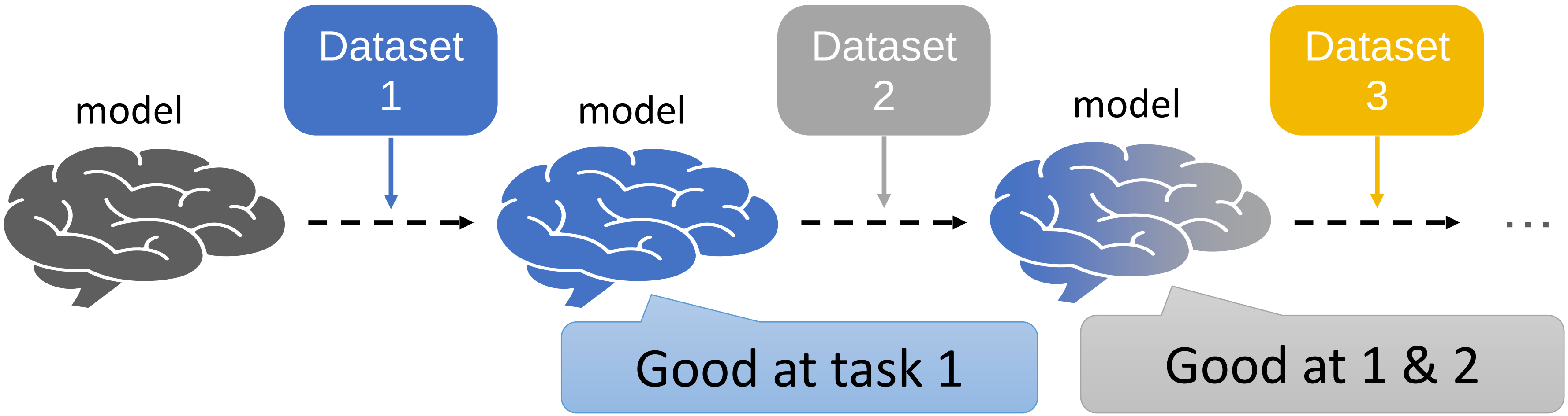



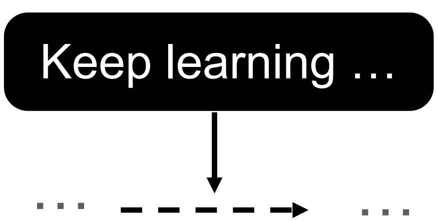

Meta learning itself also needs domain adaptation.

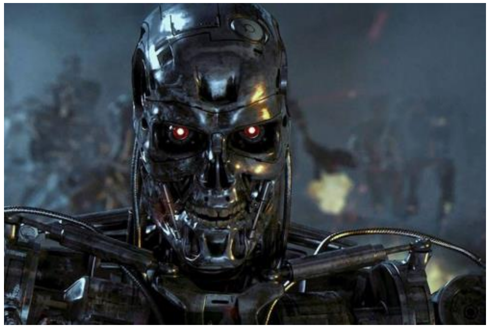
Huaxiu Yao, Longkai Huang, Linjun Zhang, Ying Wei, Li Tian, James Zou, Junzhou Huang, Zhenhui Li, Improving generalization in meta-learning via task augmentation, ICML, 2021

# Meta Learning vs. Life-long Learning

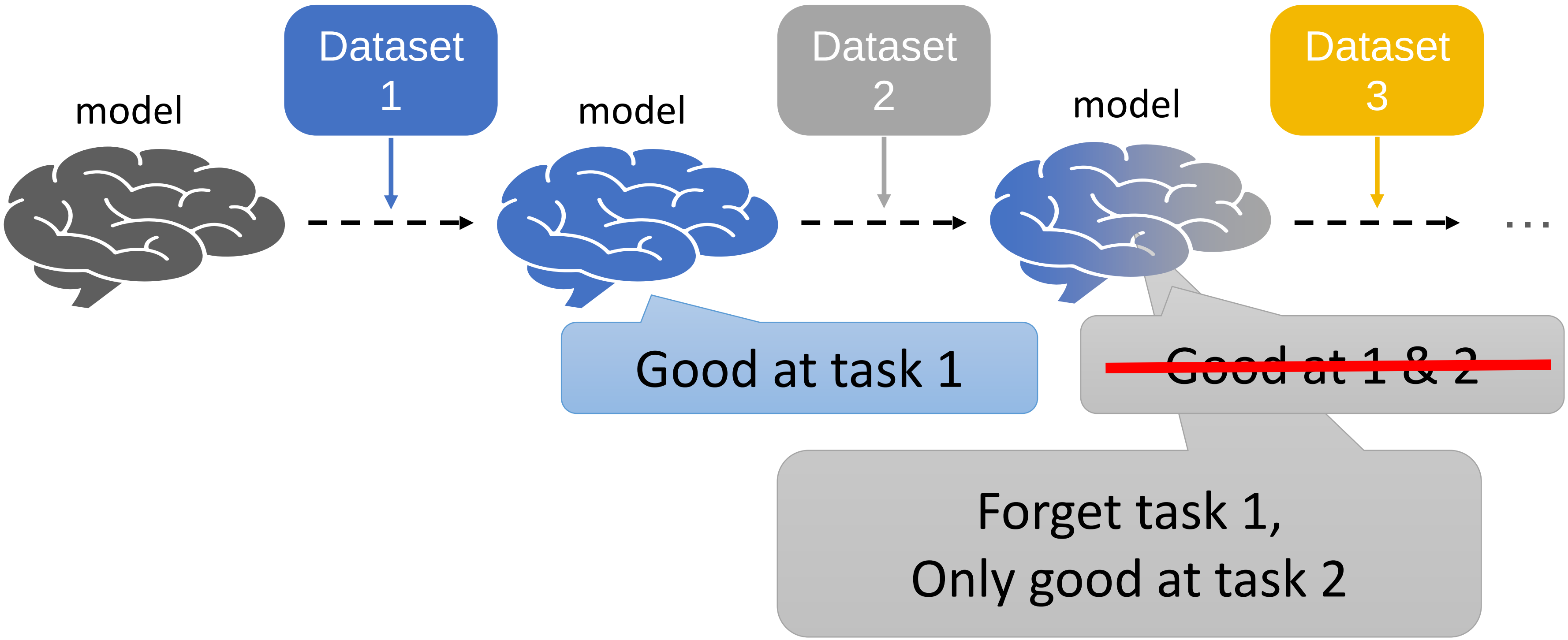
# Lifelong Learning Scenario



# Lifelong Learning Scenario




**Catastrophic forgetting!**

# Mitigating Catastrophic Forgetting

Selective Synaptic Plasticity

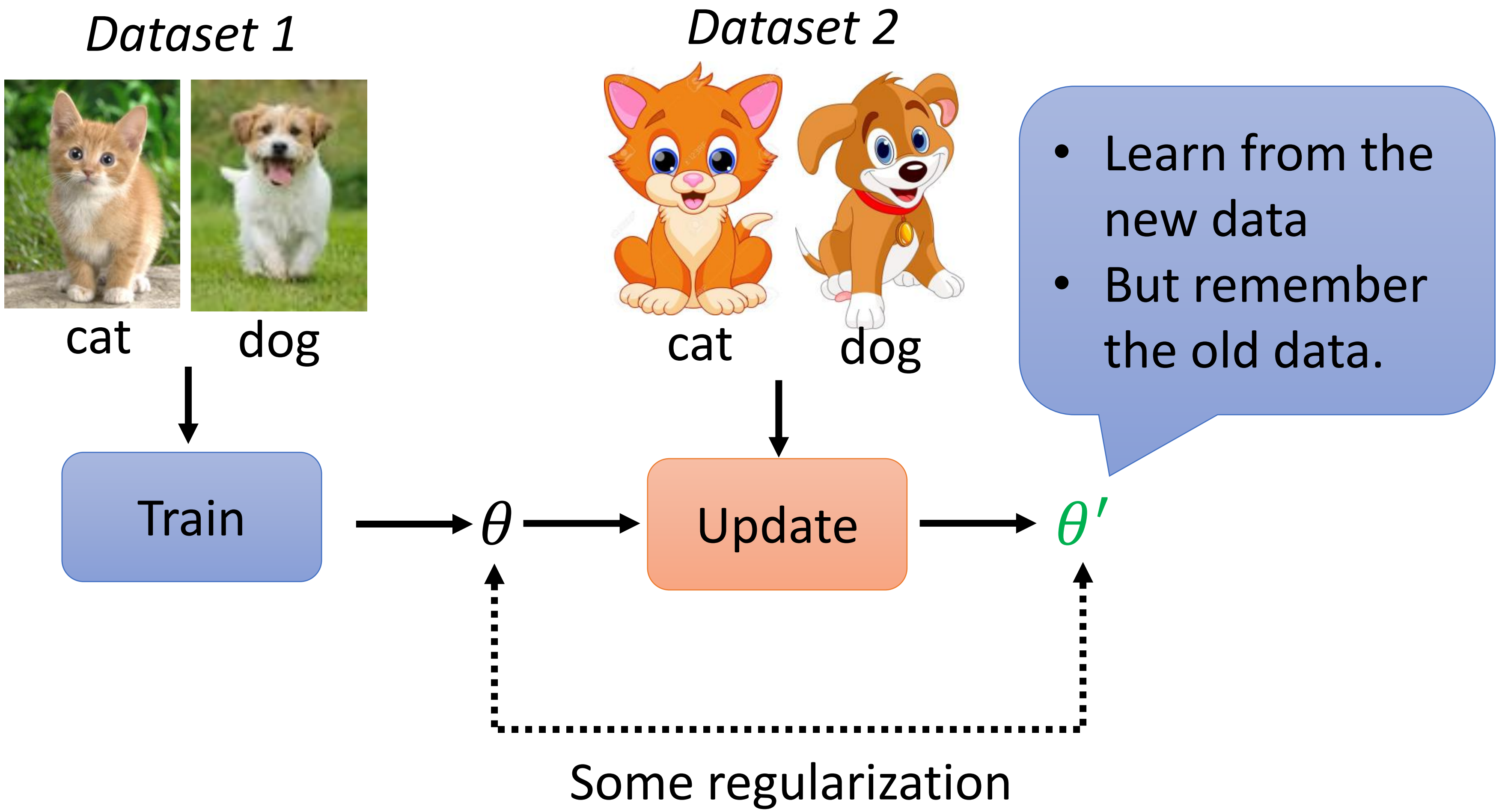
Regularization-based

Additional Neural Resource Allocation

Memory Replay

- There are already lots of research along each direction.
- Can meta learning enhance these approaches?

# Regularization-based




L2 does not work. For prevent forgetting: EWC, SI, MAS ……

# Regularization-based + Meta

*Dataset 1*

cat dog

*Dataset 2*

cat dog

- Learn from the new data
- But remember the old data.

Train → $\theta$ → Update → $\theta'$

satisfy

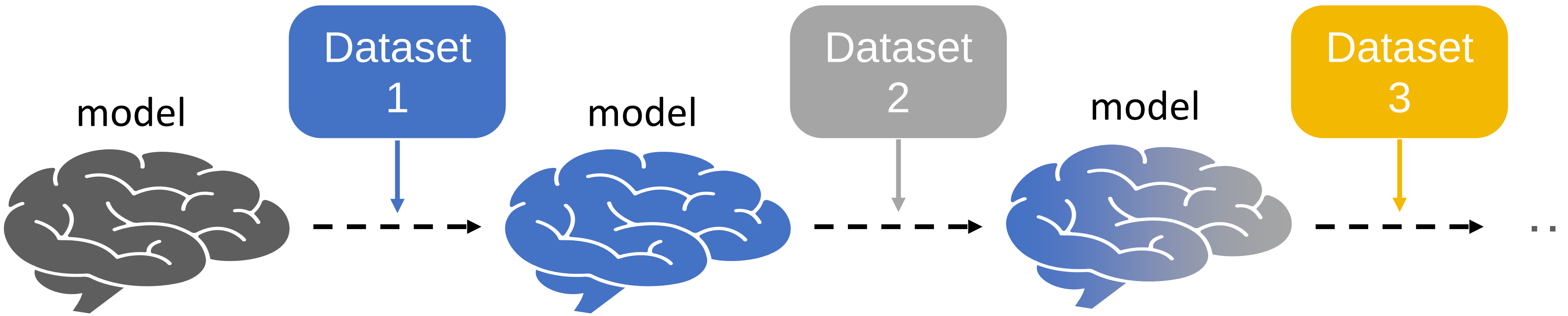
Learn by meta learning

Find a learning algorithm to prevent catastrophic forgetting.

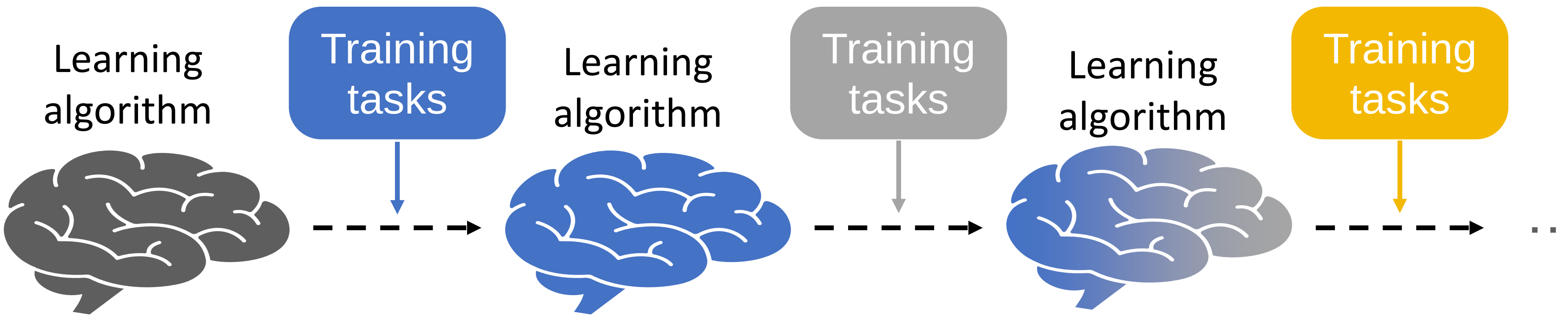
Nicola De Cao, Wilker Aziz, Ivan Titov, Editing Factual Knowledge in Language Models, EMNLP, 2021

Anton Sinitsin, Vsevolod Plokhotnyuk, Dmitriy Pyrkin, Sergei Popov, Artem Babenko, Editable Neural Networks, ICLR, 2020

# *Problem of Another Level ……*




Meta learning can help.




Meta learning itself also face the issue of catastrophic forgetting!

Chelsea Finn, Aravind Rajeswaran, Sham Kakade, Sergey Levine, Online Meta-Learning, ICML, 2019

Pauching Yap, Hippolyt Ritter, David Barber, Addressing Catastrophic Forgetting in Few-Shot Problems, ICML, 2021

# Concluding Remarks

- Meta Learning vs. Self-supervised Learning
- Meta Learning vs. Domain Generalization
- Meta Learning vs. Knowledge Distillation
- Meta Learning vs. Life-long Learning

# To Learn More ……

**Meta Learning for Natural Language Processing: A Survey**

**Hung-yi Lee**
National Taiwan University
hungyilee@ntu.edu.tw

**Shang-Wen Li**
Amazon AI*
shangwel@amazon.com

**Ngoc Thang Vu**
University of Stuttgart
thangvu@ims.uni-stuttgart.de

https://arxiv.org/abs/2205.01500

# 2022 Eighth Frederick Jelinek Memorial Summer Workshop

The application for the undergraduate research internships will be available on February 10th.  Please read the **"AI Research Internships for Undergraduates"** for more information

### The Workshop June 27 to August 5, 2022

## About the Eighth Frederick Jelinek Memorial Summer Workshop

### The JSALT 2022 Program

JHU Summer School on Human Language Technology (June 13 June 24)

Opening Day Presentations Schedule (June 27) 11:10 p.m. (GMT+8)

Plenary Lectures by Invited Speakers (June 29, July 6, 13, 20, 27)

Closing Day Presentations (August 4 and 5)

### Research Groups

- Speech Translation for Under-Resourced Languages
- Multilingual and Code-Switching Speech Recognition
- Leveraging Pre-Training Models for Speech Processing

TWS
TAIWAN
WEB
SERVICE

Thanks for supporting computing resources!